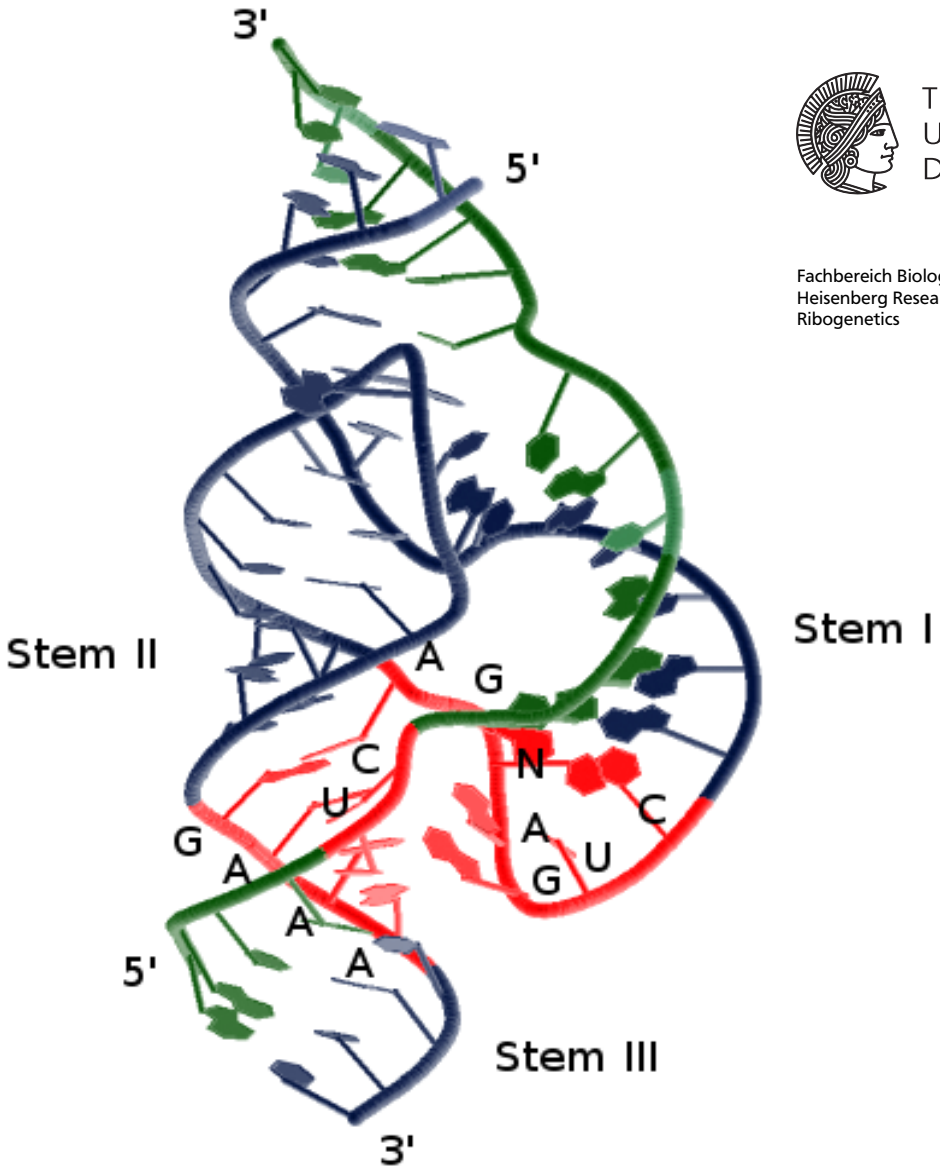
Identifikation von RNA-Motiven durch Datenbankanalysen

Identification of RNA motifs using database analysis
Zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation von MSc. Carsten Seehafer aus Cottbus
März 2013 — Darmstadt — D 17



TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich Biologie
Heisenberg Research Group
Ribogenetics

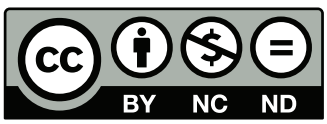
Identifikation von RNA-Motiven durch Datenbankanalysen
Identification of RNA motifs using database analysis

Genehmigte Dissertation von MSc. Carsten Seehafer aus Cottbus

1. Gutachten: Prof. Dr. Christian Hammann
2. Gutachten: Prof. Dr. Kay Hamacher

Tag der Einreichung: 18. November 2012
Tag der Prüfung: 21. März 2013

Darmstadt — D 17



Erklärung zur Dissertation

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit entsprechend den Regeln guter wissenschaftlicher Praxis selbstständig und ohne unzulässige Hilfe Dritter angefertigt habe. Sämtliche aus fremden Quellen direkt oder indirekt übernommenen Gedanken sowie sämtliche von Anderen direkt oder indirekt übernommenen Daten, Techniken und Materialien sind als solche kenntlich gemacht. Die Arbeit wurde bisher bei keiner anderen Hochschule zu Prüfungszwecken eingereicht. Während der Arbeit entstanden Publikationen, die veröffentlicht oder zur Veröffentlichung eingereicht wurden:
[Seehafer et al., 2011],
[Kalweit et al., 2011],
[Seehafer et al., 2012],
[Hoffgaard et al., Prep],
[Wiegand et al., Prep].

Darmstadt, den 24. April 2013

(Carsten Seehafer)

Kurzfassung

In dieser Dissertation wurden sequenz- und strukturbasierte Beschreibungen von RNA-Klassen, insbesondere von Ribozymen erstellt. Diese Beschreibungen dienten dazu, öffentlich zugängliche Genomdatenbanken auf Vorkommen der betreffenden Klassen zu untersuchen. Die gefundenen Motive wurden sowohl im Dateisystem, als auch in einer MySQL-Datenbank gespeichert und durch einen ProServer visualisiert. Die Arbeit zielt damit darauf ab, die Verbreitung der Motive zu analysieren und zu annotieren, um somit Einblicke in deren Evolution zu erhalten. In einem zweiten Projekt wurden RNA *Deep Sequencing* Daten eines Wildtyps und einer Gendeletionsmutante einer RNA-abhängigen RNA-Polymerase aus *Dictyostelium discoideum* ausgewertet, um den Einfluss auf Retrotransposons und die Genregulation durch RNA-Moleküle zu untersuchen.

This thesis is about the development of sequence and structure based descriptions of RNA classes, especially of ribozymes. These descriptions are used to search the classes in public available genome databases. The found motifs are stored in the file system and in a MySQL database and can be visualized by a ProServer. The aim of this thesis was the identification, analysis and annotation of the motifs and their distribution to get insights into their evolution. A second project of the thesis was the analysis of RNA Deep Sequencing data. Therefor small RNA of an RNA-dependent RNA polymerase (RdRP) wild type and a gene deletion mutant of *Dictyostelium discoideum* were analyzed to check the influence of RdRP on retrotransposable elements and gene regulation by RNA molecules.

Danksagung

Ich möchte mich bei allen bedanken, die mich während meiner Promotion unterstützt haben.

Mein besonderer und herzlicher Dank gilt meinem Doktorvater Dr. rer. nat. Christian Hammann, der durch seine Unterstützung und zahlreichen Anregungen zum Gelingen der Arbeit beigetragen hat.

Ich danke Herrn Prof. Dr. Kay Hamacher, dass er sich als Zweitgutachter zur Verfügung gestellt hat.

Des Weiteren danke ich Herrn Prof. Dr. Gerhard Steger für die gute Zusammenarbeit, für die Bereitstellung des von seinem Diplomanden modifizierten Programmes `PatScan` und weiterer Computerresourcen der Universität Düsseldorf sowie für seine Hilfe in all meinen Fragen.

Außerdem möchte ich Dr. Stefan Gräf danken, der mich unterstützte und mir viele nützliche Ideen und Hinweise gab, gerade in Fragestellungen was Ensembl betraf.

Desweiteren danke ich Dr. Johan Reimegård und Dr. Fredrik Söderbom für die Betreuung und Unterstützung bei der Auswertung der *Deep Sequencing* Daten.

Ich danke meiner Arbeitsgruppe für die angenehme Zusammenarbeit und das herzliche Klima, besonders
Anne Kalweit, die mir half viele meiner gefundenen Motive experimentell im Labor auszuwerten und somit eine verbesserte Parameterauswahl ermöglichte.

Herrn Prof. Dr. Wolfgang Nellen danke ich, dass er es mir ermöglichte innerhalb der Genetik der Universität Kassel zu arbeiten sowie der gesamten Abteilung für die gute Arbeitsatmosphäre.

Außerdem möchte ich mich bei meiner Freundin Katja Schubert bedanken für ihre Geduld und Unterstützung in dieser Zeit sowie meiner Familie.

Abkürzungsverzeichnis

API	*Application Programming Interface*
BAM	*Binary Alignment Map*
BLAST	*Basic Local Alignment Search Tool*
bzw.	beziehungsweise
CDS	*Coding Sequence*
DAS	*Distributed Annotation System*
DNA	*Deoxyribonucleic acid*
dsRNA	doppelsträngige RNA
EBI	European Bioinformatics Institute
EST	*Expressed Sequence Tag*
FP	*False Positives*
FTP	*File Transfer Protocol*
GB	Giga Byte
GFF	*General Feature Format*
GHz	Giga Hertz
glmS	*glucosamine-6-phosphate activated*
GUI	*Graphical User Interface*
HDV	Hepatitis Delta Virus
HHRz	*Hammerhead* Ribozym
HTTP	Hypertext Transfer Protokoll
IGV	*Integrative Genomics Viewer*
LINEs	*Long Interspersed Elements*
LTR	*Long Terminal Repeats*
MB	Mega Byte
MBp	Millionen Basenpaare
MHz	Mega Hertz
miRNA	*microRNA*
mRNA	*messenger RNA*
MSA	*Multiple Sequence Alignment*
NCBI	National Center for Biotechnology Information
ncRNA	*non-coding RNA*
nt	Nukleotide
ORF	*Open Reading Frames*
PCR	*Polymerase Chain Reaction*
PDB	Protein Datenbank
PHP	**P**HP: ***H**ypertext **P**reprocessor*
piRNA	*Piwi interacting RNA*
RAM	*Random Access Memory*

RdRP *RNA-dependent RNA Polymerase*
RISC *RNA-induced Silencing Complex*
RNA *Ribonucleic acid*
RNAi RNA-Interferenz
rRNA ribosomale RNA
RrpC *RNA-directed RNA Polymerase C*

SAM *Sequence Alignment Map*
SBS Sequenzierung mittels Synthese
SELEX *Systematic Evolution of Ligands by Exponential Enrichment*
SINEs *Short Interspersed Elements*
siRNA *short interfering RNA*
SLURM *Simple Linux Utility for Resource Management*
snoRNA *small nucleolar RNA*
snRNA *small nuclear RNA*
SQL *Structured Query Language*

TB Tera Byte
TDMs *Thermodynamic Matchers*
TP *True Positives*
tRNA *transfer RNA*
TSV *Tab-separated Values*

u. a. unter anderem
UCSC University of California Santa Cruz
URL *Uniform Resource Locator*
usw. und so weiter
UTR *Untranslated Region*

VS *Varkud Satellite*

WTSI Wellcome Trust Sanger Institute

XML *Extensible Markup Language*
XSL *Extensible Stylesheet Language*

z. B. zum Beispiel

Abbildungsverzeichnis

Tabellenverzeichnis

Inhaltsverzeichnis

1 Einleitung

Diese Arbeit beschäftigt sich mit der sequenz- und strukturbasierten Beschreibung und Suche von *Ribonucleic acid* (RNA)-Klassen, mit dem Ziel bekannte Klassen in Genomsequenzen zu finden und deren Verbreitung und Variation zu analysieren, um somit Einblicke in die Evolution der untersuchten Motive zu erhalten. Dabei liegt das Hauptaugenmerk auf der Klasse der Ribozyme. Diese wurden unter anderem (u. a.) in annotierten Datenbanken des Ensembl Projektes [Hubbard et al., 2009] sowie in Datenbanken des NCBI und weiteren öffentlichen Quellen gesucht (siehe Kapitel 2). Die gefundenen Treffer (siehe Kapitel 3) wurden sowohl im Dateisystem als auch in einer MySQL-Datenbank gespeichert und befinden sich zum Teil in gefilterter Form im Anhang. Diese können über einen ProServer visualisiert oder mit Hilfe von *General Feature Format* (GFF) Dateien über die Ensembl Webseite zusammen mit Annotationen dargestellt werden.

Des Weiteren wurden aus einem zweiten Projekt Illumina *Deep Sequencing* Daten aus *Dictyostelium discoideum* analysiert und annotiert.

1.1 Genetische Information

1953 zeigten James Watson und Francis Crick, dass die *Deoxyribonucleic acid* (DNA) der Träger genetischer Informationen in allen Organismen ist [Watson & Crick, 1953]. Diese Information wird laut dem zentralen Dogma der Molekularbiologie über die RNA in Proteine übertragen [Crick, 1970]. Doch die RNA ist weit mehr als ein Zwischenschritt in der Informationsübertragung. Das zentrale Dogma wurde deshalb erweitert, um darüber hinaus Prozesse und Details zu beschreiben, wie zum Beispiel (z. B.) die reverse Transkription (RNA zu DNA) [Temin, 1964] und die katalytischen Eigenschaften von RNA-Molekülen [Kruger et al., 1982]. Eine künstliche Form der Informationsspeicherung ist die *Xeno nucleic acid* (XNA), die ebenfalls der Evolution folgt und in definierte Strukturen faltet [Pinheiro et al., 2012]. Es gibt die Hypothese, dass zu Beginn der Evolution eine RNA-Welt existierte ohne Proteine. In dieser könnte die RNA gleichermaßen eine katalytische als auch eine informationsspeichernde Funktion besessen haben [Gilbert, 1986].

1.2 DNA

Die DNA besteht aus vier verschiedenen Basiseinheiten, die als Nukleotide bezeichnet werden. Jedes Nukleotid enthält ein Phosphat, einen Zucker (Desoxyribose) und eine der vier Basen: Adenin (A), Guanin (G), Cytosin (C) und Thymin (T). Adenin und Guanin sind Purine, die mit den Pyrimidinen Thymin und Cytosin Basenpaarungen eingehen. Zwischen A und T bestehen dabei zwei, zwischen G und C drei Wasserstoffbrückenbindungen. Diese werden auch als Watson-Crick Basenpaare bezeichnet [Watson & Crick, 1953]. Abbildung 1.1 zeigt den Ausschnitt der Protein Datenbank (PDB) Kristallstruktur 1ILC [Hizver et al., 2001] einer DNA. Die DNA ist eine Matrize für die *messenger RNA* (mRNA) (Transkription), welche wiederum als Vorlage für die Proteinsynthese (Translation) dient [Crick, 1970].

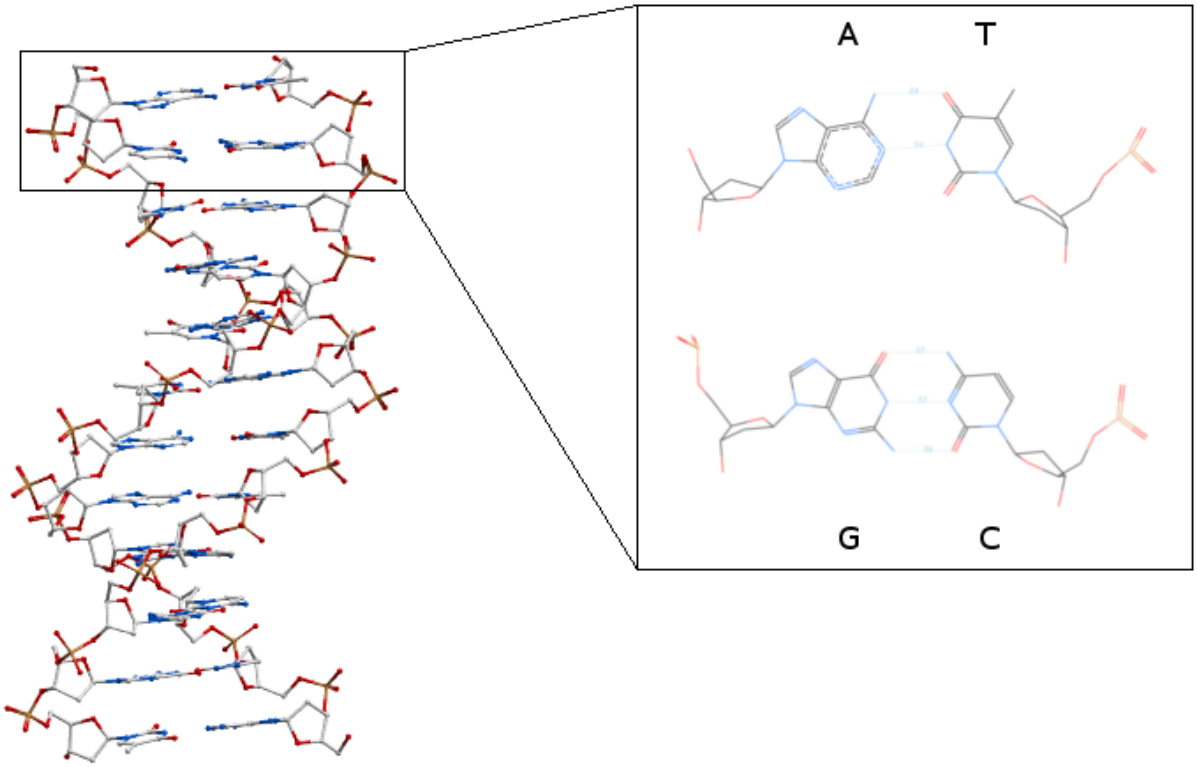


Abbildung 1.1: DNA
1ILC [Hizver et al., 2001] ist die PDB-ID einer kristallisierten DNA, als Beispiel für den Aufbau der Doppelhelix und der Watson-Crick Basenpaarungen.

1.3 RNA

Die RNA ist der DNA sehr ähnlich mit dem Unterschied, dass die Nukleotide einen anderen Zucker beinhalten. Der Zucker heißt Ribose und unterscheidet sich von der Desoxyribose in der Anzahl der Hydroxylgruppen. Ribose besitzt an der 2' Position des Pentoseringes eine Hydroxylgruppe. Außerdem existiert an Stelle der Base Thymin die Base Uracil (U). Dieser fehlt im Vergleich zu Thymin eine Methylgruppe. Neben den Watson-Crick Basenpaaren kommen in der RNA ebenfalls Wobble Basenpaare (G:U, U:G) vor [Crick, 1966]. Aufgrund unterschiedlicher räumlicher Interaktionen existieren weitere Formen der Basenpaarung. Dabei werden je nach dem, welche Kante die Wasserstoffbrückenbindungen ausbildet drei Kanten unterschieden: Watson-Crick-Kante, Hoogsteen-Kante und Zucker-Kante [Leontis & Westhof, 2001] (Abbildung 1.2). Außerdem wird anhand der relativen Orientierung der glykosidischen Bindungen ein Basenpaar als *cis* oder *trans* eingeordnet [Leontis & Westhof, 2001]. Es gibt doppel- und einzelsträngige RNA mit zahlreichen Funktionen. Je nach Funktion gibt es verschiedene Arten, die im Abschnitt 1.3.2 erläutert werden. Die Funktionalität ergibt sich durch die optimale thermodynamische Faltung der Sequenz. Es wird dabei zwischen Primär-, Sekundär-, Tertiärstruktur unterschieden.

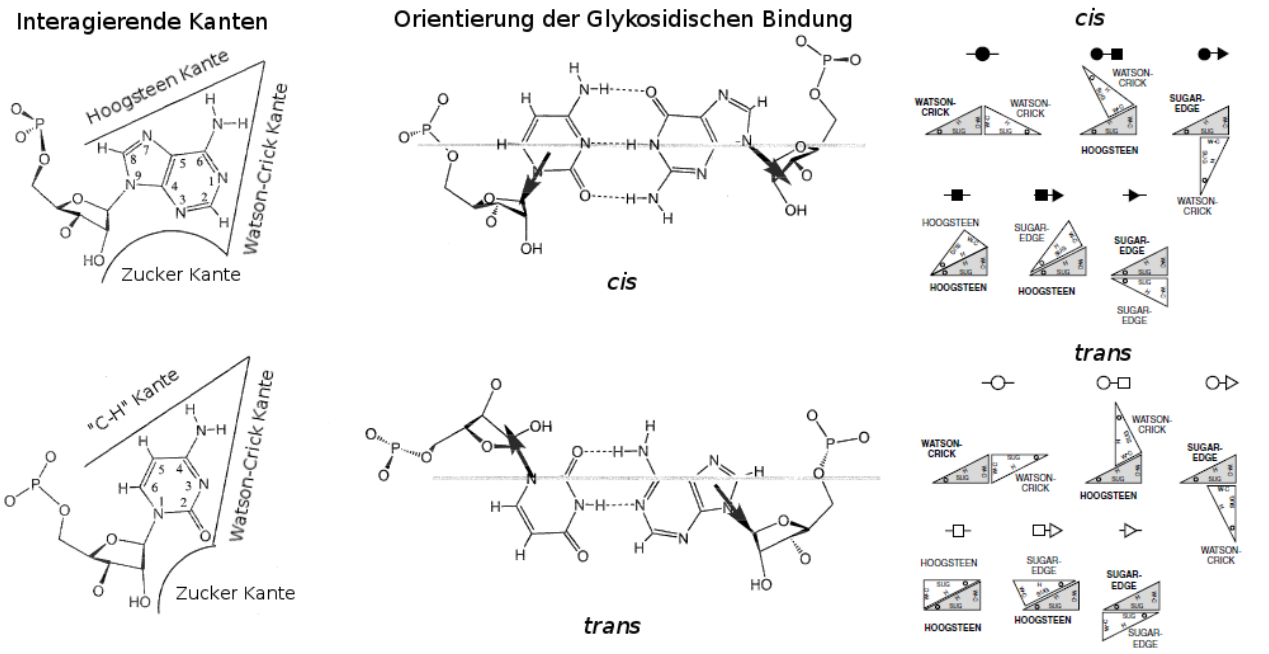


Abbildung 1.2: RNA-Interaktion
Dargestellt sind drei Interaktionsmöglichkeiten von Adenosin und Cytosin sowie die Orientierung der glykosidischen Bindung, angepasst aus [Leontis & Westhof, 2001, Leontis et al., 2002].

1.3.1 Strukturen

Die Primärstruktur der RNA ist die Betrachtung der Struktur auf Sequenzebene. Die einzelnen Nukleotide sind durch Phosphodiesterbindungen miteinander verbunden. Zur Beschreibung der Primärstruktur werden IUPAC-Symbole [Cornish-Bowden, 1985] verwendet (Tabelle 1.1).

[1]

Tabelle 1.1: IUPAC-Symbole

Symbol	Bedeutung	Symbol	Bedeutung
A	A	K	G oder U
C	C	V	nicht U
G	G	H	nicht G
U oder T	U	D	nicht C
M	A oder C	B	nicht A
R	A oder G	N	A,C,G,U
W	A oder U	X	mögliches A,C,G,U
S	C oder G	.	nichts
Y	C oder U	-	gap

[1] Nomenklatur für unvollständig spezifizierte Basen in Nukleinsäuresequenzen. X wurde für diese Arbeit hinzugefügt und steht für eine mögliche zusätzliche Position mit Symbol N.

Als Sekundärstruktur wird eine Struktur bezeichnet, die sich aus Wasserstoffbrücken und Stapelwechselwirkungen der Basen zusammensetzt. Die Definition der Strukturen wurden aus [Zuker & Stiegler, 1981] übernommen. Eine solche Struktur ist z. B. der *Hairpin Loop*, der aus einem Basenpaar und einem *Loop* besteht. Der *Loop* setzt sich aus ungepaarten Basen zusammen. Ein *Bulge-Loop* besitzt ungepaarte Basen auf einer Seite des Doppelstranges, wodurch die Struktur gebogen wird. Bei einem *Internal-Loop* sind beide Seiten ungepaart. Werden mehrere Helices miteinander verbunden, wird dies als *Junction, Bifurcation* oder *Multi-Loop* bezeichnet [Zuker & Stiegler, 1981, Dirks et al., 2004]. Im Vergleich zu Helices sind *Loops* energetisch ungünstig und führen zur Destabilisierung der RNA [Zuker & Stiegler, 1981]. Die Sekundärstruktur kann auf verschiedene Weisen dargestellt werden. Eine Darstellungsweise ist z. B. die Punkt-Klammer-Notation, bei der Punkte ungepaarte Basen symbolisieren und öffnende mit zugehörigen schließenden Klammern Basenpaarungen bedeuten. Eine andere Darstellungsmöglichkeit ist die Verwendung gleichwinkliger Polygone, wie in Abbildung 1.3 zu sehen [Hofacker et al., 1994].

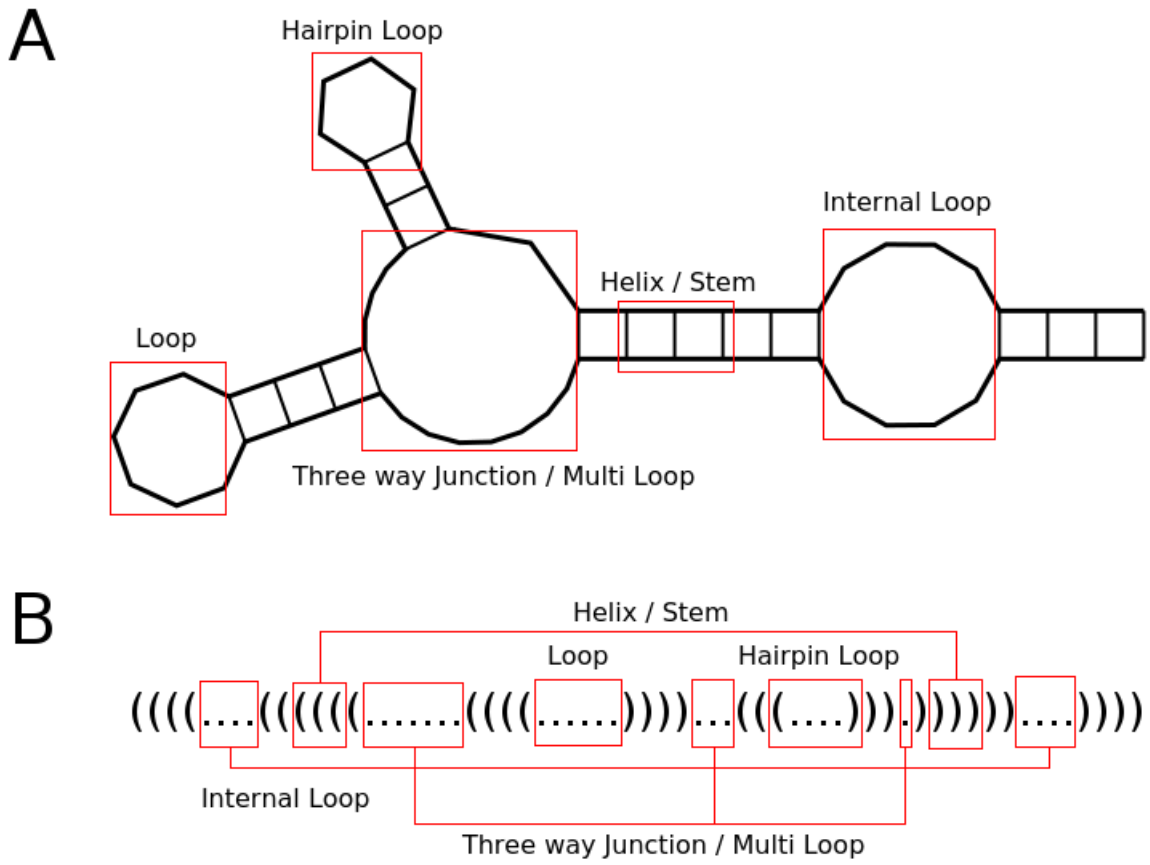


Abbildung 1.3: RNA-Sekundärstrukturelemente
A) Polygonale Darstellung und als B) Punkt-Klammer-Struktur von PDB-ID 2GOZ [Martick & Scott, 2006] erstellt mit RNAfold und KolourPaint.

Die Tertiärstruktur setzt sich aus den einzelnen Sekundärstrukturelementen in einer Kette zusammen, die durch Wasserstoffbrückenbindungen und Stapelwechselwirkungen stabilisiert werden. Sie beschreibt ebenfalls die Interaktion der einzelnen Elemente [Moss, 1996]. Tertiärstrukturen sind z. B. *Tetraloop-Tetraloop* Rezeptorinteraktionen, Pseudoknoten, Ribose Zipper oder coaxiale helikale Stapelung [Reiter et al., 2011].

1.3.2 Arten

Die RNA zeigt eine große Vielfalt in Anzahl, Sequenz, Größe, Struktur und Funktion, aufgrund dessen wird sie in mehrere Arten untergliedert. An dieser Stelle sei nur eine Auswahl von RNA näher vorgestellt.

messenger RNA - Die mRNA ist der Informationsträger für die Translation [Crick, 1970]. Aus eukaryontischer *precursor*-mRNA (pre-mRNA) werden beim *Splicing* Introns entfernt (zusammengefasst in [Wahl et al., 2009]). Die resultierende mRNA besteht von 5' nach 3' aus einem *Cap*, der 5' *Untranslated Region* (UTR), einer *Coding Sequence* (CDS), der 3' UTR und einer Poly(A)-Kette. Die mRNA kann auf Sekundärstrukturebene zum Teil *cis* agierende RNA-Elemente beinhalten [McGuire & Galagan, 2008].

transfer RNA - Die *transfer RNA* (tRNA) ist der Träger einer spezifischen Aminosäure und bindet während der Translation an das entsprechende Basentriplet (Codon) auf der mRNA [Crick, 1968].

ribosomale RNA - Die ribosomale RNA (rRNA) ist ein Bestandteil der Ribosomen und beinhaltet die enzymatische Aktivität für die Translation [Nissen et al., 2000].

small nuclear RNA - sind kleine nukleare RNA, die im Zellkern, während des *Splicings*, Introns aus der pre-mRNA herausschneiden [Hüttenhofer et al., 2001].

small nucleolar RNA - *small nucleolar RNA* (snoRNA) sind kleine nukleoläre RNA, die je nach Sequenzmotiv in Box C/D und Box H/ACA Subklassen unterteilt werden. Diese sind nach der Transkription für die Methylierung und Pseudouridylierung von rRNA und *small nuclear RNA* (snRNA) verantwortlich [Hüttenhofer et al., 2001]. Durch die zahlreichen RNA-Ziele werden weitere zelluläre Funktionen vermutet [Bachellerie et al., 2002]. Die meisten snoRNA in Säugetieren und Pflanzen befinden sich in Introns Protein-kodierender Gene [Hertel et al., 2007] und in intergenischen Regionen [Jöchl et al., 2008].

microRNA - Diese in Viren, einzelligen und mehrzelligen Organismen vorkommenden, Genom-kodierten RNA werden aus einem gewöhnlich nicht perfekt doppelsträngigen *hairpin precursor* durch Dicer Proteine prozessiert und ergeben doppelsträngige RNA (dsRNA) mit einer Länge von $\sim$ 21 Nukleotiden (nt) [Li & Ding, 2005, Grivna et al., 2006, Hinas et al., 2007]. Nach der Auftrennung durch eine Helikase in zwei Einzelstränge gehen sie mit *RNA-induced Silencing Complex* (RISC) einen Komplex ein und binden komplementär an die mRNA, so dass diese nicht translatiert oder gespalten wird [Bartel, 2004].

short interfering RNA - Diese kleinen RNA sind das Produkt der Prozessierung langer doppelsträngiger RNA durch Dicer Proteine in $\sim$ 21 nt lange Fragmente (primäre siRNA), welche mit RISC einen Komplex eingehen [Tomari et al., 2004]. Die dsRNA kann ihren Ursprung aus Viren oder repetitiven Elementen haben oder durch eine *RNA-dependent RNA Polymerase* (RdRP) aus einer einzelsträngigen RNA synthetisiert worden sein (sekundäre siRNA) [Hinas et al., 2007]. Die *short interfering RNA* (siRNA) kann somit den Abbau von Fremd-RNA initialisieren [Bartel, 2004].

Piwi interacting RNA - sind kleine RNA, die zusammen mit Piwi Proteinen einen Komplex eingehen. Sie besitzen eine Länge von 26-31 nt und werden durch eine Methylierung am 3' Ende stabilisiert. Sie kommen sowohl im Cytoplasma als auch im Zellkern vor. Eine Ziel-Sequenz wird nach dem zehnten Nukleotid gespalten. Da *Piwi interacting RNA* (piRNA) komplementär zu transposablen Elementen sind, wird angenommen, dass sie an der Hemmung von Retrotransposons beteiligt sind. Dafür spricht ebenfalls die Beobachtung, dass das Entfernen von piwi Proteinen zu einem Anstieg der Transposons führt [Aravin et al., 2008].

non-coding RNA - Zelluläre RNA, die keine Funktion als mRNA, tRNA oder rRNA besitzen, aber auf RNA-Ebene von Bedeutung sind, werden als nicht Protein kodierende RNA (ncRNA) bezeichnet [Hüttenhofer & Vogel, 2006, Jöchl et al., 2008]. Sie sind kleiner als 500 nt und somit kürzer als die meisten mRNA, wobei es jedoch Ausnahmen gibt, z. B. Xist RNA [Hong et al., 1999] oder Air RNA [Sleutels et al., 2002]. Genomweite Suchen in verschiedenen Modellorganismen zeigten unerwartet viele *non-coding RNA* (ncRNA) in allen Domänen des Lebens, von Archaeen über Bakterien bis hin zu Eukaryonten [Hüttenhofer & Vogel, 2006]. In Bakterien wird funktionale ncRNA meist in intergenischen Bereichen kodiert [Hüttenhofer & Vogel, 2006] .

Die Funktionen zellulärer RNA sind ebenso vielfältig, wie ihre Arten. Dazu zählen u. a. regulative, modifizierende, stabilisierende und transportierende Prozesse [Hüttenhofer & Vogel, 2006, Jöchl et al., 2008]. Manche Funktionen sind noch immer nicht entschlüsselt.

1.3.3 Retrotransposons

Retrotransposons, als Untergruppe der Transposons (*class* I), sind mobile Elemente, die über einen RNA-Zwischenschritt vervielfältigt und in ein Genom eingefügt werden können [Deininger & Batzer, 2002]. Es werden dabei zwei Typen, mit und ohne *Long Terminal Repeats* (LTR), unterschieden. LTR-Retrotransposons ähneln Retroviren und besitzen transkriptionsregulatorische Sequenzen in den LTR und verschiedene *Open Reading Frames* (ORF), jedoch ohne Verpackungsgene [Deininger & Batzer, 2002]. Retrotransposons können für Proteine kodieren, die essentiell für die Transposition sind, wie z. B. für eine Endonuklease oder eine reverse Transkriptase, mit der aus RNA DNA-Elemente erzeugt werden [Mathias et al., 1991]. Ein Beispiel für ein LTR-Retrotransposon ist DIRS-1. Es ist in der Amöbe *Dictyostelium discoideum* das häufigste, mobile Element [Eichinger et al., 2005], mit einer linken und invertierten rechten LTR, drei ORF und einer Gesamtlänge von ∼4500 Basenpaaren [Glöckner et al., 2001]. Ein anderes häufig vorkommendes LTR-Retrotransposon ist Skipper. Non-LTR Retrotransposons werden in *Short Interspersed Elements* (SINEs) und *Long Interspersed Elements* (LINEs) unterteilt [Deininger & Batzer, 2002]. SINEs sind kurze Elemente, die keine Protein-kodierenden Gene besitzen und durch RNA-Polymerase III transkribiert werden. Sie besitzen am 3' Ende eine Poly(A)-Kette und sind von der Transpositionsmaschinerie anderer mobiler Elemente abhängig. LINEs dagegen sind unabhängig. Fast alle mobilen Elemente sind an der Integrationsstelle von kurzen *Repeats* umgeben [Deininger & Batzer, 2002]. Bei Insertion in Genen können Pseudogene entstehen, die ihre Funktion verloren haben [Vanin, 1985]. Transposon Sprünge zählen zu den Hauptgründen für genomische Instabilität [Vastenhouw et al., 2003]. Repetitive Elemente, wie Transposons sind in vielen Eukaryonten Bestandteil von Centromeren [Hinas et al., 2007].

1.3.4 Katalytisch aktive RNA

RNA-Moleküle, die eine chemische Reaktion katalysieren, werden als Ribozyme bezeichnet [Kruger et al., 1982]. Katalytisch aktive RNA wurde erstmals in den 1980er Jahren in Gruppe I Introns [Cech et al., 1981] und bakterieller Ribonuklease P [Guerrier-Takada et al., 1983] beschrieben. Gruppe I Introns, Gruppe II Introns und Ribonuklease P sind größer als 200 Nukleotide (nt), bestehen aus mehreren Domänen und zählen zu den großen Ribozymen. Ihre aktiven Zentren werden durch mehrere *Junctions*, Helices und Einzelstrangregionen gebildet [Golden, 2011]. Die Ribosomen, bestehend aus ribosomaler RNA und Protein, sind Orte der Proteinsynthese und können als Ribozyme bezeichnet werden, da Kristallstrukturen im katalytischen Zentrum ausschließlich RNA gezeigt haben [Nissen et al., 2000]. Die kleinen Ribozyme (50-150nt) [Hammann & Lilley, 2002] sind besonders für die Untersuchung von RNA-Struktur-Funktionen geeignet [Birikh et al., 1997, Perreault et al., 2011]. Zu diesen zählen vier strukturelle Klassen selbst-spaltender RNA:

- *Hammerhead* Ribozyme [Prody et al., 1986],
- *Hairpin* Ribozyme [Buzayan et al., 1986a, Rupert & Ferré-D'Amaré, 2001],
- Hepatitis Delta Virus (HDV) Ribozyme [Kuo et al., 1988, Wu et al., 1989, Ferré-D'Amaré et al., 1998],
- *Varkud Satellite* (VS) Ribozym [Saville & Collins, 1990, Lilley, 2004]

[Hammann & Westhof, 2007, Cochrane & Strobel, 2008].

Ribozyme sind an vielen fundamentalen biologischen Prozessen beteiligt, wie z. B. beim *Splicing*, bei der Translation, bei der Genregulation und der RNA-Prozessierung [Golden, 2011]. Sie werden in der Biotechnologie als Werkzeug eingesetzt, um RNA-Sequenzen zu spalten [Seemann & Hartig, 2011, Giliberti et al., 2012].

Chemische Reaktion

Alle Mitglieder dieser vier strukturellen Klassen vollziehen eine endonukleolytische Spaltung beziehungsweise (bzw.) Ligation des Phosphodiesterrückgrates der eigenen Sequenz [Cochrane & Strobel, 2008, Fedor, 2009]. Dabei übt, im Fall einer Spaltung, die 2' Hydroxylgruppe einen nukleophilen Angriff auf die 3',5' Phosphodiesterbindung aus. In dem anschließenden trigonalen, bipyramidalen Übergangszustand erfolgt eine nukleophile Substitution (S_N2-Reaktion genannt), mit dem Ergebnis, dass sich beim 5' Produkt am Ende ein 2', 3' zyklisches Phosphat und beim 3' Produkt eine 5' Hydroxylgruppe bildet (siehe Abbildung 1.4 A). Diese können Substrate für die reverse Ligationsreaktion sein, die durch den gleichen Übergangszustand führt [Wilson & Lilley, 2009]. Die Reaktion wird durch unterschiedliche Elemente der RNA und / oder zweiwertige Metallionen erreicht. Die Spaltung bzw. Ligation kann in *trans* oder *cis* erfolgen (siehe Abbildung 1.4 B, C am Beispiel des *Hammerhead* Ribozyms) [Luzi et al., 1997, Martick et al., 2008].

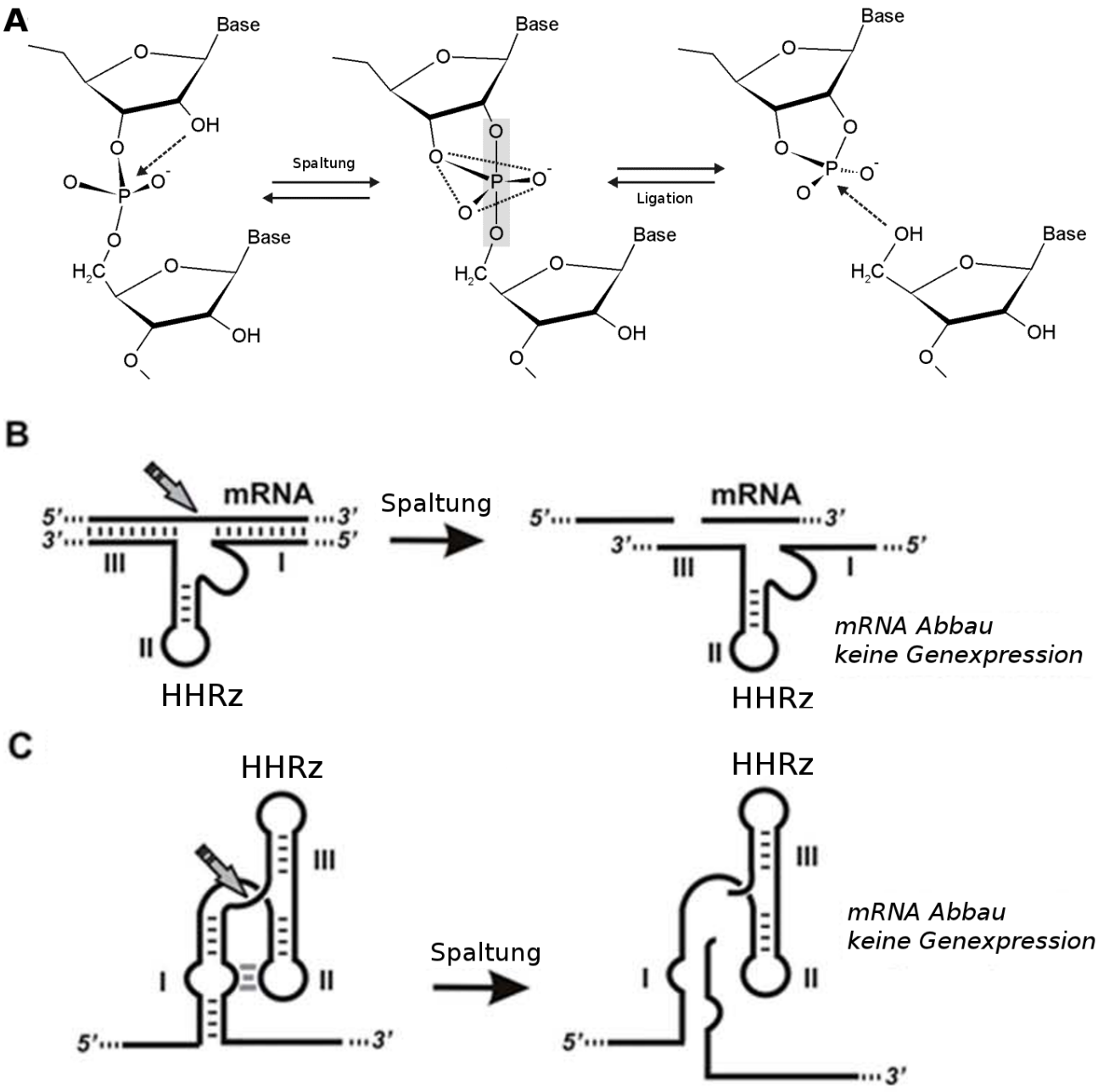


Abbildung 1.4: S_N2-Mechanismus
A) zeigt eine Spaltungs- bzw. Ligationsreaktion, bei der sich im Fall einer Spaltung 5' ein 2', 3' zyklisches Phosphat und 3' eine 5' Hydroxylgruppe bildet (angepasst aus [Kalweit et al., 2011]). B) und C) wurden modifiziert aus [Seemann & Hartig, 2011] übernommen und zeigen die Spaltung eines *Hammerhead* Ribozyms in *trans* und *cis*.

Hammerhead Ribozyme sind kleine *non-coding RNA* (ncRNA), die eine selbst-spaltende bzw. selbst-ligierende Aktivität besitzen. Sie bestehen aus drei Helices, die durch Einzelstrangregionen miteinander verbunden sind und eine *three way junction* ergeben. Die Sekundärstruktur ähnelt dabei der eines Hammerkopfes, der diesen Ribozymen den Namen verleiht. Die Helices und Einzelstrangregionen besitzen variable Längen. Helix II und III bestehen zum Teil aus nur ein bis zwei Basenpaaren [Perreault et al., 2011] und *Loop* I kann hunderte Nukleotide groß sein [Martick et al., 2008]. Nur das katalytische Zentrum bestehend aus 11 konservierten Nukleotiden [Uhlenbeck, 1987] ist konstant, wobei mittlerweile mehrere neue strukturelle Variationen gezeigt werden konnten (zusammengefasst in [Hammann et al., 2012]). Zur einheitlichen Bezeichnung der Positionen wurden die Nukleotide nach [Hertel et al., 1992] nummeriert. Eine solche strukturelle Variation ist z. B. eine Insertion vor C3, die in Phagen gefunden wurde, welche in einer sehr salzhaltigen Umgebung vorkommen [Perreault et al., 2011]. Je nachdem, welche Helix das offene 5', 3' Ende besitzt, werden 3 Typen von *Hammerhead* Ribozymen unterschieden. Alle 3 Typen sind in Abbildung 1.5 [Hoffgaard et al., Prep] zu sehen.

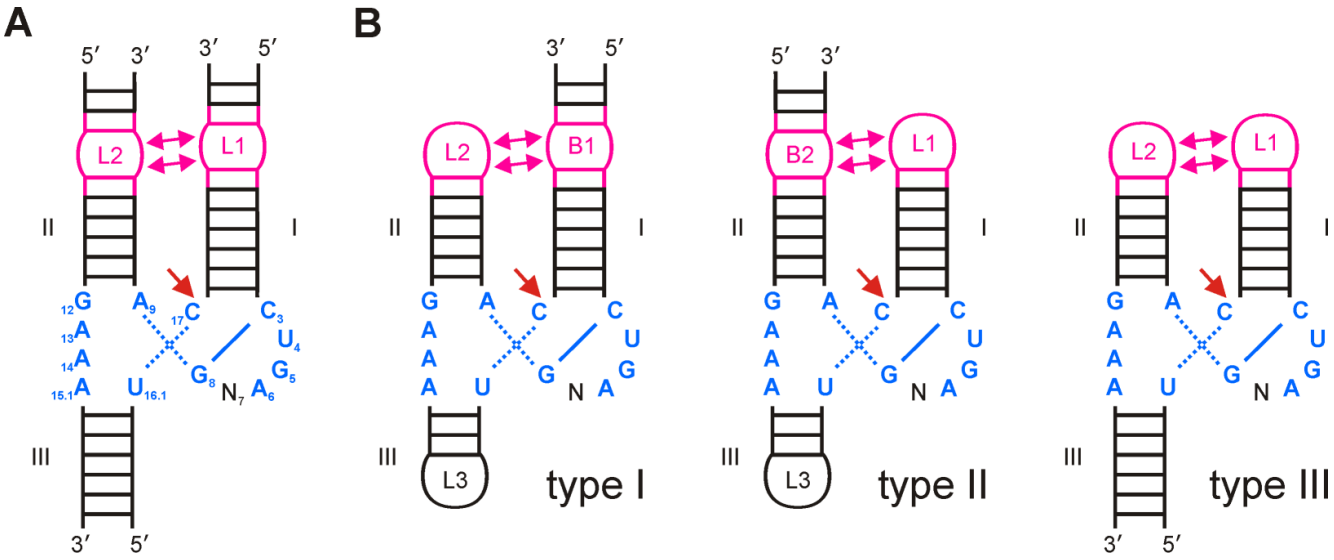
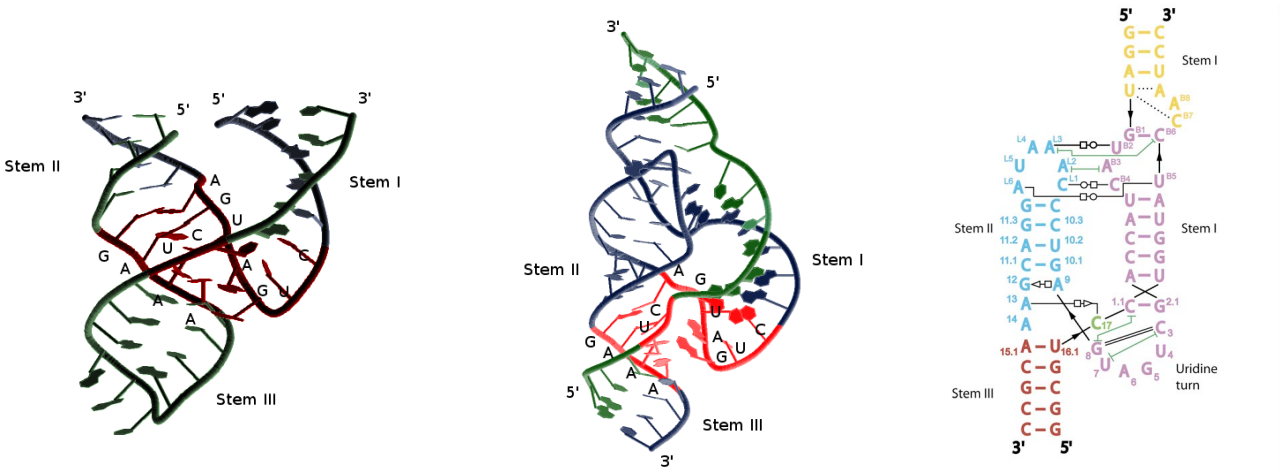


Abbildung 1.5: *Hammerhead* Ribozym Typen
A) zeigt die allgemeine Struktur eines *Hammerhead* Ribozyms mit den konservierten Nukleotiden des katalytischen Zentrums (blau) und der mit dem roten Pfeil markierten Spaltstelle. Die konservierten Nukleotide sind nach [Hertel et al., 1992] durchnummeriert. B) stellt die drei möglichen *Hammerhead* Ribozym (HHRz) Typen dar, je nach geöffneter Helix.

Typ I und III kommen in der Natur vor. 1988 zeigten Haseloff und Gerlach, dass durch Öffnen des *Stem-Loop* II ein in *trans* spaltendes *Hammerhead* Ribozym vom Typ II möglich ist [Haseloff & Gerlach, 1988], aber bis 2011 konnte kein natürlich existierendes Typ II *Hammerhead* Ribozym (HHRz) nachgewiesen werden. Jimenez *et al.* zeigte schließlich, dass es sich bei dem gefundenen HHRz II sogar um das bis zu diesem Zeitpunkt schnellste bekannte *cis* spaltende HHRz handelte [Jimenez et al., 2011]. Inzwischen ist bekannt, dass Typ II HHRz in Bakterien genauso häufig zu finden sind, wie Typ I und III [Perreault et al., 2011].

Die minimal Struktur des HHRz spaltet nur bei einer sehr hohen Mg^{2+} Konzentration. Diese wird durch periphäre Elemente drastisch reduziert, da zwischen den Nukleotiden der *Loops* L1 und L2 bzw. zwischen dem internen *Loop* B1 und L2 (Typ I) oder L1 und B2 (Typ II) sowie den Helices I und II in der Tertiärstruktur Interaktionen bestehen [Canny et al., 2004, Osborne et al., 2005, Martick & Scott, 2006, Roychowdhury-Saha et al., 2011]. Die Interaktionen führen zu einer veränderten Metallionenbindung des katalytischen Zentrums und somit zu einem Anstieg der Spaltungsgeschwindigkeit [de la Peña et al., 2003, Khvorova et al., 2003, Penedo et al., 2004, Canny et al., 2004]. Allgemein handelt es sich um nicht kanonische Basenpaarungen und Stapelung individueller, entfernter Basen, die nur eine geringe Konservierung auf der Sequenz- und Strukturebene aufweisen [Martick & Scott, 2006, Przybilski & Hammann, 2006, Chi et al., 2008, Dufour et al., 2009]. In 40% der gefundenen HHRz-Kandidaten von Perreault *et al.* sind diese Interaktionen Pseudoknoten [Perreault et al., 2011]. Nur wenn die Interaktionen stattfinden, kann eine katalytische Aktivität unter physiologischen Bedingungen beobachtet werden [de la Peña et al., 2003, Khvorova et al., 2003, Martick & Scott, 2006]. Dazu zählt eine der Zelle entsprechende Mg^{2+} Konzentration im millimolaren Bereich. Alternativ könnten auch andere divalente Ionen, wie Fe^{2+}, verwendet werden [Athavale et al., 2012].



(a) 1MME (b) 2GOZ (c) 2GOZ

Abbildung 1.6: *Hammerhead* Ribozym Typ I
(a) zeigt in 2 Ketten die Kristallstrukturen eines minimalen [Scott et al., 1995] und (b) eines erweiterten *Hammerhead* Ribozyms [Martick & Scott, 2006] erstellt mit Jmol und Gimp. (c) entnommen aus [Martick & Scott, 2006] zeigt die Tertiärstruktur von 2GOZ und besitzt zusätzliche periphäre Elemente, die miteinander interagieren und durch die Leontis-Westhof Nomenklatur [Leontis & Westhof, 2001] dargestellt sind.

Auch die Bildung des N3N8 Basenpaares ist neben dem pH-Wert und G5, A14 für eine katalytische Aktivität entscheidend [Ruffner et al., 1990, Przybilski & Hammann, 2007, Buskiewicz & Burke, 2012]. Abbildung 1.6 zeigt das Titelbild der Arbeit, mit der ersten Kristallstruktur eines vollständigen *Hammerhead* Ribozyms vom Typ I [Martick & Scott, 2006].

Hammerhead Ribozyme wurden erstmals 1986 in Viroiden [Hutchins et al., 1986] und in Satelliten RNA des *Tobacco ringspot virus* [Prody et al., 1986] entdeckt. Beide RNA-Moleküle sind zirkulär und einzelsträngig mit zahlreichen intramolekularen Basenpaarungen und einer Länge von 250-400 nt [Buzayan et al., 1986b, Tabler & Tsagris, 2004]. Der Unterschied zwischen Viroiden und Satelliten RNA liegt in der Replikation. Satelliten RNA benötigen einen Hilfsvirus. Viroide können sich autonom vervielfältigen, besitzen kein Kapsid und kodieren für keine Proteine [Elena et al., 1991].

Es wurden seit der ersten Entdeckung zahlreiche Exemplare des HHRz gefunden, wie z. B. von

- Epstein und Gall in repetitiver satelliten DNA des Molches (*Triturus carnifex*) [Epstein & Gall, 1987] und in weiteren Amphibien [Zhang & Epstein, 1996],
- Rojas *et al*. in Höhlenschrecken (*Dolichopoda*) [Rojas et al., 2000],
- Ferbeyre *et al*. in Pärchenegel *Schistosoma mansoni* [Ferbeyre et al., 1998],
- Daròs und Flores im *Sense*- und *Antisense* Strang von DNA Tandem *Repeats* der Nelke (*Dianthus caryophyllus*) [Daròs & Flores, 1995], was auf einen genomischen Einbau Viroid ähnlicher Sequenzen deutet,
- Przybilski *et al*. in der Ackerschmalwand (*Arabidopsis thaliana*) auf Chromosom 4 [Przybilski et al., 2005],
- Martick *et al*. in verschiedenen anderen Eukaryonten [Martick et al., 2008].

Erste Datenbanksuchen gaben bereits Hinweise auf eine breite Verteilung der HHRz, obwohl die theoretische Wahrscheinlichkeit gering ist [Ferbeyre et al., 2000].

In repetitiven Regionen können HHRz in Tandem hintereinander vorkommen [Perreault et al., 2011]. Für eine bessere Stabilisierung sind ebenfalls verschachtelte, z. B. über Stem III miteinander verbundene Formen möglich, wie künstlich in der Publikation [Forster & Symons, 1987] gezeigt wurde.

HHRz dienen in sub-viralen Systemen der Herstellung einheitlich langer Fragmente für die Replikation, die nach dem symmetrischen oder asymmetrischen *„Rolling circle“* Mechanismus abläuft [Hammann & Steger, 2012].

Eingebettet in der 5'UTR bakterieller mRNA regulieren natürlich vorkommende HHRz die Genexpression von Stoffwechselgenen [Martick et al., 2008]. Dies konnte ebenfalls für eukaryotische mRNA gezeigt werden, wo sie z. B. eingebettet in der 3'UTR von *C-Typ Lektin Typ II* (*Clec2*) mRNA, durch Selbstspaltung, posttranskriptional die Genexpression kontrollieren [Martick et al., 2008]. Bei einer inaktiven Mutante wurde die mRNA nicht gespalten und ergab ein intaktes 3'UTR Transkript. Das Gen hat einen Einfluss auf die Knochenremodellierung und Immunantwort von Säugetieren [Martick et al., 2008].

Das genomische Umfeld der neu gefundenen Sequenzen in Pro- und Eukaryonten lässt auf weitere biologische Funktionen schließen [Perreault et al., 2011]. In vielen Fällen ist die biologische Funktion jedoch unklar. *Hammerhead* Ribozyme werden in verschiedenen Geweben exprimiert und treten *in vivo* in gespaltener und ungespaltener Form auf [Przybilski et al., 2005].

Es wird angenommen, dass sie sich unabhängig voneinander in den verschiedenen Organismen entwickelt [Salehi-Ashtiani & Szostak, 2001] und wahrscheinlich über mobile Elemente verbreitet haben [Przybilski et al., 2005].

Hairpin Ribozyme

Ein weiteres selbst-spaltendes RNA-Motiv ist das *Hairpin* Ribozym [Rupert & Ferré-D'Amaré, 2001]. Das *Hairpin* Ribozym besteht in zwei viralen Satelliten RNA aus einer *four way junction* und in einer weiteren aus einer *five way junction* [Bajaj et al., 2011]. Das minimale *Hairpin* Ribozym setzt sich aus zwei Helix-Internal *Loop*-Helix Segmenten zusammen, wobei die *Loops* in der aktiv gefalteten Struktur miteinander interagieren, um die Spaltung zu erleichtern [Fedor, 1999]. Dazu müssen strukturelle Umordnungen erfolgen [Suydam et al., 2010].

Die Interaktion wird durch divalente Metallionen, ein Watson-Crick GC Basenpaar und durch Ribose Zipper aus Nukleotiden der *Loops* stabilisiert [Rupert & Ferré-D'Amaré, 2001].

Abbildung 1.7 zeigt bekannte Sequenzen in denen *Hairpin* Ribozyme gefunden wurden.

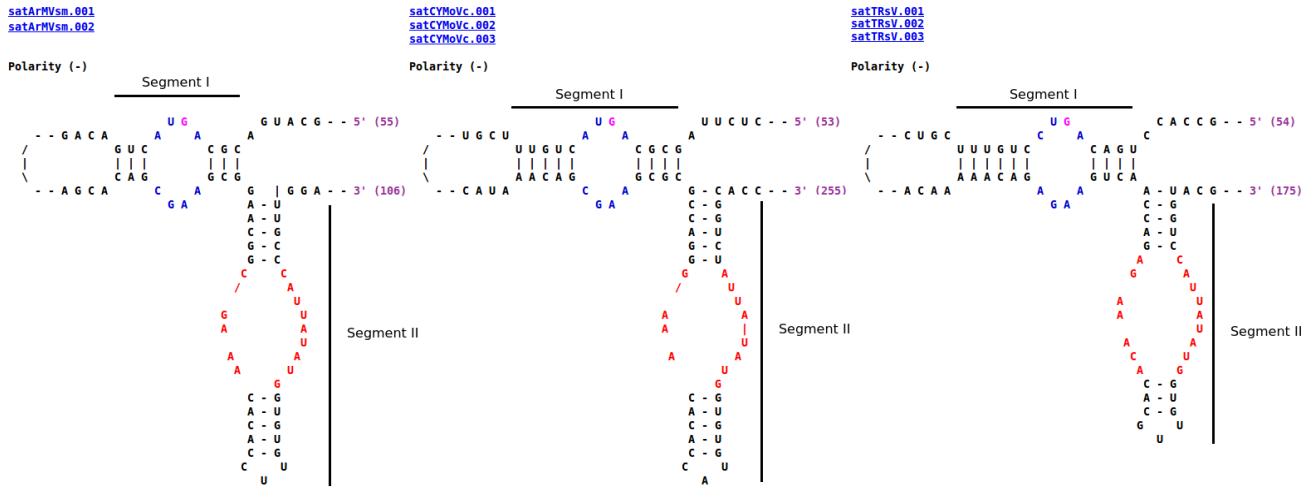


Abbildung 1.7: *Hairpin* Ribozyme
Bekannte Sequenzen aus *Arabis mosaic virus small satellite RNA* (satArMVsm) [Kaper et al., 1988], *Chicory yellow mottle virus satellite RNA* (satCYMoVc) [Rubino et al., 1990] und *Tobacco ringspot virus satellite RNA* (satTRsV) [Buzayan et al., 1986b] aus SubviralDB [Rocheleau & Pelchat, 2006]. Die zwei Helix-Internal *Loop*-Helix Segmente sind markiert. Segment I enthält die Spaltstelle.

Daraus ergibt sich für das Segment II die folgende Konsensussequenz:

`NNNNN NGAAA NNNN X CNU X NNNN GUAUAUUAN NNNNN`

In den in der Legende zu Abbildung 1.7 aufgelisteten Viren dienen die *Hairpin* Ribozyme der Prozessierung von Satelliten RNA Transkripten, die während der *Rolling circle* Replikation entstehen.

Die von Rupert *et al.* generierte Kristallstruktur eines *Hairpin* Ribozyms in einer katalytisch aktiven Konformation ist in Abbildung 1.8 zu sehen [Rupert & Ferré-D'Amaré, 2001].

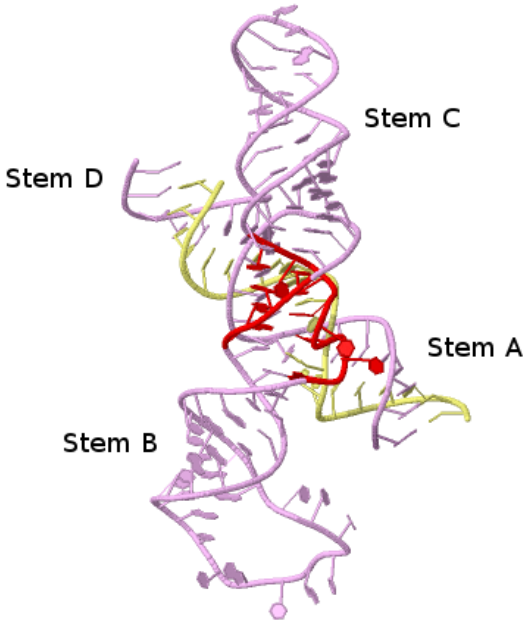


Abbildung 1.8: *Hairpin* Ribozym Kristallstruktur
1M5K [Rupert & Ferré-D'Amaré, 2001] ist die PDB-ID eines kristallisierten, katalytisch aktiven *Hairpin* Ribozyms. Dargestellt sind zwei Ketten, in gelb das RNA-Substrat und in lila das *Hairpin* Ribozym. In rot markiert ist der *internal Loop* B (GAAA und GUAUAUUA), ähnlich wie in Abbildung 1.7. 1.8 wurde mit Jmol erzeugt und Gimp beschriftet.

Hairpin Ribozyme bevorzugen die Ligationsreaktion, wobei im minimalen Konstrukt auch die Spaltungsreaktion stattfinden kann [Fedor, 1999]. Sie sind die ersten Ribozyme, die für Gentherapien eingesetzt wurden [Shippy et al., 1999].

Hepatitis Delta Virus Ribozyme

Der Hepatitis Delta Virus (HDV) besteht aus einer zirkulären, viroid-ähnlichen RNA, die eine Ribozymaktivität besitzt, welche während der *Rolling circle* Replikation von großer Bedeutung für den Virus ist [Lai, 1995]. Das HDV Ribozym setzt sich aus einem kompakten, katalytischen Zentrum mit fünf helikalen Segmenten zusammen, die über zwei verschachtelte Pseudoknoten miteinander verbunden sind (Abbildung 1.9) [Ferré-D'Amaré et al., 1998].

Das HDV Ribozym verwendet zwei unterschiedliche katalytische Strategien für die Spaltungsreaktion. Zum Einen eine Säure-Base Reaktion, in der das Cytosin (C75) als Protonendonor fungiert und zum Anderen eine Metallionen-Reaktion, bei der Mg^{2+} Ionen mit der 2' Hydroxylgruppe der Ribose von U(-1) und der Phosphatgruppe interagieren [Golden, 2011].

Eine Ligationsreaktion der HDV Ribozyme konnte bislang nicht beobachtet werden [Chen et al., 2009].

Die Analyse HDV ähnlicher Ribozyme im Moskito hat gezeigt, dass die Spaltungsrate in den verschiedenen Entwicklungsstadien variiert, was auf eine Regulation der Ribozyme schließen lässt [Webb et al., 2009].

Ruminski *et al.* konnten zeigen, dass verschiedene Retrotransposons aus Arthropoden, Nematoden und Chordata mit Hilfe von HDV ähnlichen Ribozymen am 5' Ende gespalten werden. Sie steuern außerdem *in vitro* und *in vivo* die Translationinitiation von *downstream* liegenden ORF [Ruminski et al., 2011].

HDV Ribozyme werden ebenfalls als Werkzeug für die Gentherapie eingesetzt [Lévesque & Perreault, 2012].

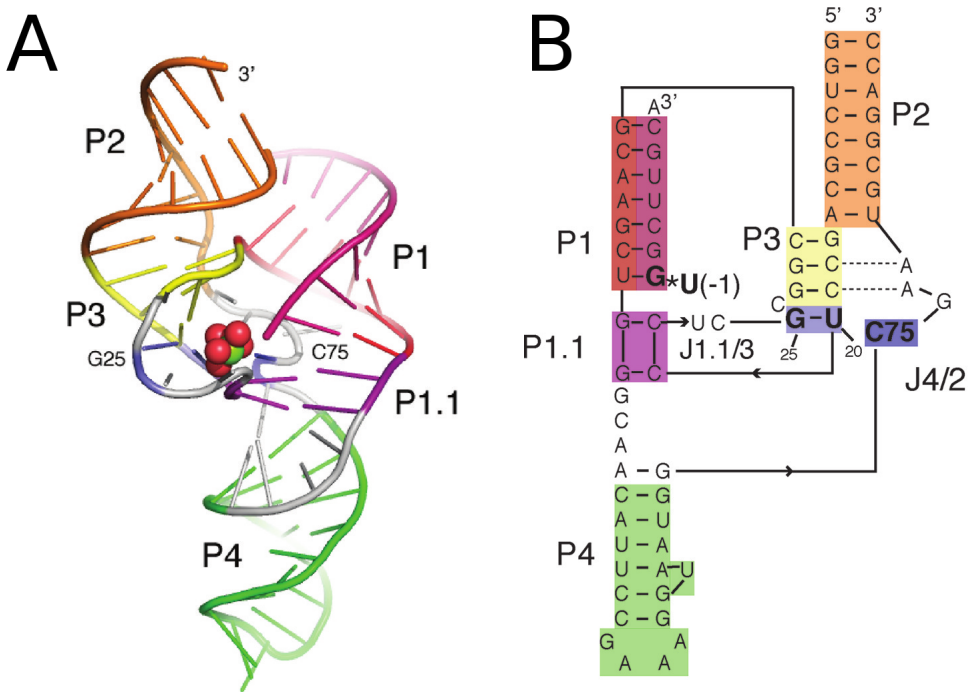


Abbildung 1.9: Hepatitis Delta Virus Ribozym Kristallstruktur
A) zeigt die Kristallstruktur eines HDV Ribozyms vor der Spaltung in *trans* (PDB:3NKB). Die Spaltstelle befindet sich zwischen den Helices P1, P1.1 und P3 und ist in der Sekundärstruktur (B) mit einem Stern markiert. Das katalytische Zentrum besteht aus fünf Helices (P1-P4 und P1.1). Das Substrat (U(-1) bis A8) bildet mit den komplementären Basen des Ribozyms die Helix P1. J1.1/3 und J4/2 sind *Junctions*, die zur Spaltstelle gehören. 1.9 wurde aus [Golden, 2011], ursprünglich aus [Chen et al., 2010], modifiziert übernommen.

Varkud Satellite Ribozym

Das *Neurospora Varkud Satellite* (VS) Ribozym [Saville & Collins, 1990] ist bisher das einzige Beispiel dieses Motivs und wurde in Transkripten mitochondrieller DNA gefunden. Es befindet sich in einer 881 nt langen, einzelsträngigen, zirkulären RNA, die mit HDV, Gruppe I Introns, Retrotransposons und Satelliten RNA Gemeinsamkeiten besitzt. Die Spaltungsreaktion erzeugt eine typische 5' Hydroxyl- und 2', 3' zyklische Phosphatgruppe [Saville & Collins, 1990]. Die Spaltungs- und Ligationsreaktion findet im *internal Loop* der Helix I statt. Das gesamte Ribozym besteht aus sieben helikalen Segmenten, die durch drei *three way junctions* miteinander verbunden sind (Abbildung 1.10) [Wilson & Lilley, 2011]. Des Weiteren besteht zwischen Helix I und Helix V eine Interaktion [Rastogi et al., 1996]. Helix I befindet sich wahrscheinlich zwischen den Helices II und VI [Lafontaine et al., 2002]. Außerdem besitzt der *internal Loop* der Helix VI einen Einfluss auf die Spaltung (besonders A756), da dieser wahrscheinlich mit dem *internal Loop* der Helix I (G638) das aktive Zentrum bildet [Lipfert et al., 2008].

Die Spaltung könnte in *Neurospora* der Prozessierung multimerer RNA-Fragmente in der Replikation dienen [Kennell et al., 1995]. Es gibt zwar keine strukturelle Ähnlichkeit zu anderen Ribozymen, aber das katalytische Zentrum besitzt eine große Ähnlichkeit zu *Hairpin* Ribozymen [Wilson & Lilley, 2011].

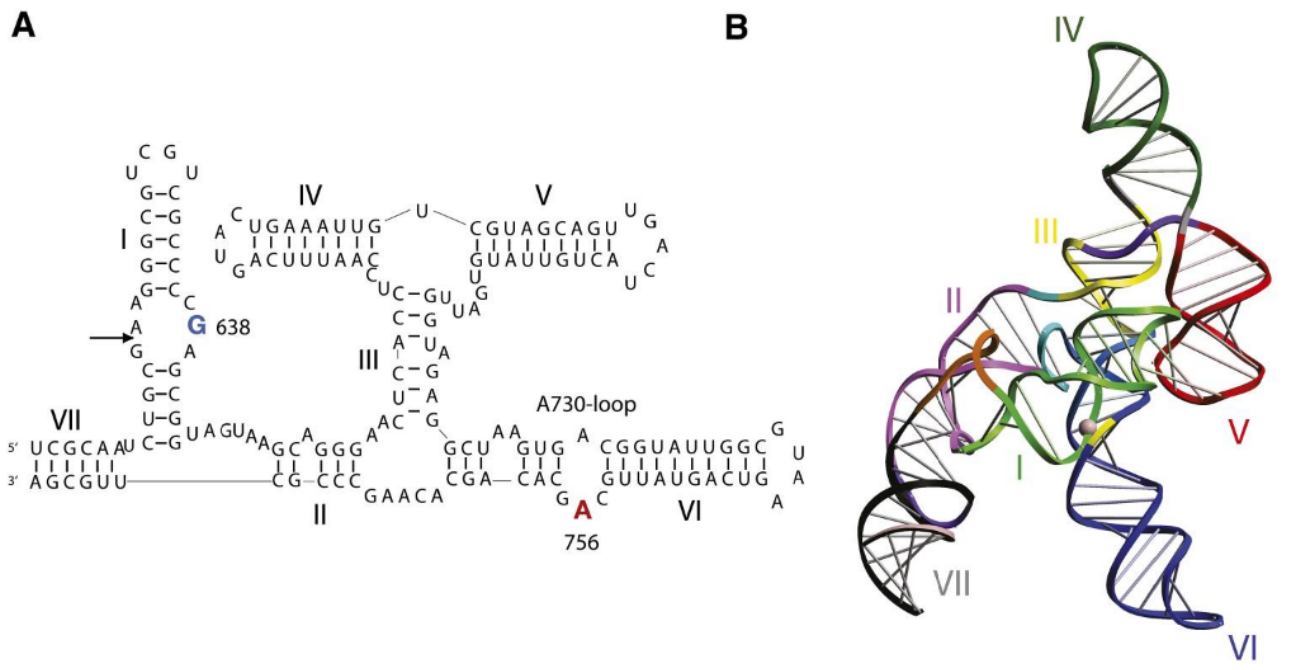


Abbildung 1.10: *Neurospora* Varkud Satellite Ribozym
A) zeigt die Sekundärstruktur des VS Ribozyms mit den von I bis VII gekennzeichneten Helices. Die Spaltstelle befindet sich im *internal Loop* der Helix I und ist mit dem Pfeil markiert. Für die Spaltung wird der A730 *Loop* in Helix VI benötigt. G638 und A756 sind besonders hervorgehoben. In B) ist das Modell einer 3D Struktur zu sehen [Lipfert et al., 2008], da diese bis jetzt nicht kristallisiert werden konnte. Das wahrscheinlich aktive Zentrum (G638 und A756) ist gelb markiert [Lipfert et al., 2008]. 1.10 wurde aus [Wilson & Lilley, 2011] übernommen und modifiziert.

Riboswitches

Riboswitches sind regulatorische Motive, die durch die hochspezifische Bindung kleiner Moleküle an- bzw. ausgeschalten werden und somit die Genregulation beeinflussen können. Bestehend aus einem Aptamer und einer Expressionsplattform kontrollieren sie essentielle Gene in vielen Bakterien durch z. B. Transkriptionstermination oder Translationsinitiation [Deigan & Ferré-D'Amaré, 2011]. Alle bekannten eukaryontischen (Thiaminpyrophosphat) *Riboswitches* haben durch *Splicing* einen Einfluss auf die Genexpression [Deigan & Ferré-D'Amaré, 2011]. Insgesamt sind zur Zeit 17 *Riboswitch*-Klassen bekannt [Breaker, 2012]. Unter allen *Riboswitches* ist das *glucosamine-6-phosphate activated* (glmS) *Riboswitch* das einzige Ribozym. Es kommt in der 5'UTR der mRNA von glmS Genen in Gram-positiven Bakterien vor und kann sich selbst in Anwesenheit von Glukosamin-6-Phosphat spalten, so dass mit steigender Glukosamin-6-Phosphat Konzentration die Anzahl der glmS Transkripte gesenkt wird [Winkler et al., 2004, Klein & Ferré-D'Amaré, 2006, Viladoms et al., 2011]. Es existieren zahlreiche künstliche *Riboswitches*, die als Werkzeug zur Genregulation eingesetzt werden [Deigan & Ferré-D'Amaré, 2011].

Rolling circle Replikation

Die *Rolling circle* Replikation (Abbildung 1.11) beschreibt die symmetrische Replikation von Satelliten RNA [Prody et al., 1986]. Die Replikation besteht aus drei Stufen: der RNA-Vervielfältigung, Spaltung und Ligation. Die zirkuläre, einzelsträngige (+) RNA dient als Matrize der Synthese multimerer (-) RNA, welche durch Selbstspaltung (*Hairpin* oder *Hammerhead* Ribozym) zu monomeren Einheiten prozessiert werden. Nach der Ligation dient die entstandene zirkuläre (-) RNA als Matrize für die Synthese multimerer (+) RNA, welche durch darin enthaltene HHRz-Motive erneut in monomere Einheiten gespalten werden [Prody et al., 1986, Feldstein et al., 1989]. Obwohl beide Ribozyme *in vitro* ligieren können, ist die Ligation der Satelliten RNA zu zirkulären Monomeren *in vivo* nicht bewiesen.

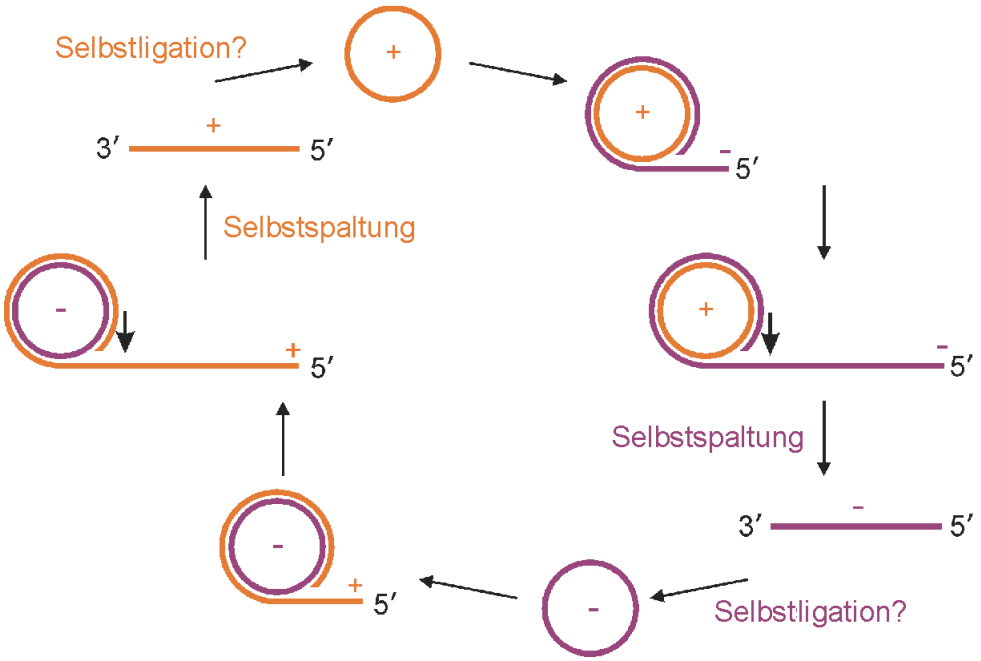


Abbildung 1.11: *Rolling circle* Replikation

1.11 zeigt das Modell der *Rolling circle* Replikation [Branch & Robertson, 1984]. Die (+) RNA dient als Vorlage für die Synthese multimerer (-) RNA, die durch Selbstspaltung der Ribozyme zu monomeren Einheiten verarbeitet werden und wiederum nach der Ligation als Vorlage für die Synthese multimerer (+) RNA dienen. 1.11 wurden aus [Seehafer et al., 2012] übernommen und modifiziert.

1.3.5 RNA-Interferenz

Der RNA-Interferenz (RNAi) Mechanismus spielt eine wichtige Rolle in der Genregulation, der Chromatin-Modifikation und in der Abwehr parasitärer Gene einschließlich Transposons und Viren [Zhang & Ruvkun, 2012]. Seit der Entdeckung des RNAi-Mechanismus 1998 in *Caenorhabditis elegans* [Fire et al., 1998] ist bekannt, dass RNAi in Eukaryonten stattfindet. Fire *et al.* injizierten eine doppelsträngige RNA (dsRNA) in eine Zelle, was zum Abbau der komplementären mRNA führte. Viele Eukaryonten verwenden für den RNAi-Mechanismus *RNA-dependent RNA Polymerases*, um dsRNA zu synthetisieren, welche daraufhin durch die Endoribonuklease Dicer zu kleinen RNA (*microRNA* (miRNA), *short interfering RNA* (siRNA)) prozessiert werden, die *RNA-induced Silencing Complex* (RISC) zu einer Ziel RNA führen [Bernstein et al., 2001], diese komplementär binden und spalten [Maida & Masutomi, 2011].

RNA-dependent RNA Polymerase

RNA-dependent RNA Polymerase (RdRP) wird im RNAi-Mechanismus verwendet und ist an verschiedenen Funktionen wie der Genregulation [Dunoyer et al., 2010], der RNA gesteuerten DNA-Methylierung [Wassenegger et al., 1994, Molnar et al., 2010] und in *Schizosaccharomyces pombe* an der Heterochromatinbildung beteiligt [Motamedi et al., 2004, Verdel et al., 2004]. RdRP produzieren komplementäre Sequenzen für eine Ziel RNA über einen Dicer-abhängigen Mechanismus oder über einen Dicer-unabhängigen Mechanismus mit Hilfe von *Argonaute*-Proteinen, die Bestandteil von RISC sind [Maida & Masutomi, 2011]. Ein alternativer RNAi-Mechanismus ist die Verlängerung der als Primer dienenden primären siRNA durch RdRP und Erzeugung doppelsträngiger RNA, die dann wiederum durch Dicer prozessiert wird und sekundäre siRNA erzeugt [Martens et al., 2002]. In *Caenorhabditis elegans* sind die Mehrheit, der über die RdRP (z. B. RRF-1 oder EGO-1) produzierten siRNA, sekundäre siRNA [Pak & Fire, 2007]. Dabei ist EGO-1 für die posttranskriptionale Regulation von Keimbahn spezifischen Genen verantwortlich [Maniar & Fire, 2011] und RRF-1 für die Regulation zellspezifischer Gene [Gent et al., 2010]. RRF-1 ist jedoch nicht essentiell [Gent et al., 2009]. Sekundäre siRNA werden ebenso in *Arabidopsis thaliana* über die RdRP RDR6 produziert, selbst für Sequenzen außerhalb der primären Zielregion, was als transitives *Silencing* bezeichnet wird [Moissiard et al., 2007].

RdRP sind für Viren von zentraler Bedeutung und dienen der Replikation und Transkription des Genoms während einer Infektion [Verdaguer & Ferrer-Orta, 2012]. Genom-kodierte RdRP wurden bislang nur in Pilzen, Pflanzen und Würmern gefunden [Li & Ding, 2005]. In Säugetieren, speziell im Menschen wurde ein Proteinkomplex (hTERT) mit einer RdRP Aktivität gefunden, dessen synthetisierte dsRNA zu siRNA prozessiert werden [Maida et al., 2009]. Die RdRP (QDE-1) aus *Neurospora crassa* kann sowohl Primer-unabhängig kleine (9-21 nt) RNA als auch Primer-abhängig lange *Antisense* RNA aus einer einzelsträngigen Vorlage synthetisieren [Makeyev & Bamford, 2002]. In *Dictyostelium discoideum* gibt es drei Gene (*rrpA, rrpB, rrpC*), die für RdRP kodieren [Martens et al., 2002]. Diese besitzen eine N-Terminale Helikase Domäne, die in anderen Organismen Teil der Dicer sind [Martens et al., 2002]. RrpA und RrpB besitzen auf der Gesamtlänge eine hohe Sequenzähnlichkeit von 97%, während die Ähnlichkeit zu *RNA-directed RNA Polymerase C* (RrpC) auf 65% Länge nur 35% beträgt [Martens et al., 2002]. Von Stephan Wiegand durchgeführte Northern Blots in einem *rrpC*-Gendeletionsstamm zeigten eine große Reduzierung kleiner RNA, vor allem für das Retrotransposon DIRS-1 [Wiegand et al., Prep]. Dagegen ist die Menge kleiner RNA für Skipper stark erhöht [Wiegand et al., Prep]. Des Weiteren wurde in Gendeletionsstämmen ohne RrpC eine erhöhte Anzahl miRNA gefunden [Hinas et al., 2007]. Im Vergleich dazu gibt es auf Transkript Ebene im *rrpC*-Gendeletionsstamm eine erhöhte mRNA Expression für DIRS-1 [Kuhlmann et al., 2005].

1.4 Motivsuche

Die Identifizierung von unbekannten Motiven aus biologischen Sequenzen ist bekannt als das „Motiv-Such-Problem“ [Rajasekaran et al., 2005]. Das Problem besteht darin, aus mehreren gegebenen Sequenzen ein unbekanntes Motiv zu identifizieren, das an verschiedenen unbekannten Positionen auftreten kann, vorausgesetzt die Sequenzen enthalten ein gemeinsames Motiv. Da Sequenzen Mutationen unterliegen, ist das Zählen von *Substring*-Häufigkeiten als Motiverkennung nicht geeignet, weshalb Substitutionen, Insertionen und Deletionen bei der Suche beachtet werden müssen. Die wahrscheinlich beste Methode, kurze Signale (<10 nt) zu finden, ist *Pattern* basierte Algorithmen zu nutzen, die jedes *Pattern* mit einem Score bewerten und somit *High-Scoring Patterns* bestimmen können [Brazma et al., 1998]. Ein weiteres Problem ist, dass viele biologische Sequenzen unvollständig sind und nur teilweise das gesuchte Signal enthalten. Das Ziel der Suche unbekannter Motive ist die Lokalisierung regulatorischer Bereiche und die Identifizierung möglicher Wirkstoffbindungsstellen [Pevzner & Sze, 2000].

Bekannte Motive können im Gegensatz dazu durch die Eingabe von Sequenz- und Strukturinformationen mit Hilfe einer Motivbeschreibungssprache durch den Benutzer definiert werden [Gräf et al., 2005]. Dies geschieht am besten in Kombination mit einer phylogenetischen Analyse [Rivas et al., 2001]. Die Beschreibung ist in schriftlicher, wie z. B. bei RNAmotif [Macke, 2001] oder in graphischer Form, wie bei Locomotif [Reeder et al., 2007] möglich. Andere Suchprogramme wiederum sind speziell auf ein bestimmtes Motiv ausgelegt, wie z. B. tRNAscan-SE [Lowe & Eddy, 1997]. Es besteht auch die Möglichkeit zuerst Sekundärstrukturelemente aus homologen Sequenzen generieren zu lassen, z. B. mit CMfinder [Yao et al., 2006], und anschließend diese mit einem *Pattern*-Suchprogramm zu durchsuchen.

De la Peña und Garcia-Robles verfolgten in ihrer Motivsuche den Ansatz, zunächst nach kurzen Teilstücken des Motivs zu suchen, bevor diese bei vorhandener, korrekter Sekundärstruktur erweitert und anschließend erneut auf korrekte Strukturinformation überprüft wurden [de la Peña & Garcia-Robles, 2010b]. Mit RMdetect ist es möglich, aus einer Kristallstrukturinformation und einem *Multiple Sequence Alignment* (MSA) ein Modell zu erzeugen, dass ohne weitere Sequenzinformation ein homologes Motiv in einer Sequenz finden kann [Cruz & Westhof, 2011]. Eine alternative ähnliche Suchform sind *Thermodynamic Matchers* (TDMs) [Reeder & Giegerich, 2004], die unabhängig von einer Motivbeschreibungssprache etablierte thermodynamische Modelle verwenden, um die Sequenz darauf zu überprüfen, ob sie sich in ein vorgegebenes Motiv faltet. Der Nachteil dieser Methode ist, dass für jede Struktur ein neues Faltungsprogramm geschrieben werden muss [Reeder et al., 2007].

Die Identifizierung von ncRNA ist schwieriger als für Protein-kodierende mRNA, da ncRNA keine typischen Sequenzeigenschaften besitzen, wie Splice- oder Translationssignale oder lange ORF [Babak et al., 2005].

RNA-Motive und ihre Funktionalität sind mit wenig Sequenzähnlichkeit hauptsächlich auf Sekundär- und Tertiärstrukturebene konserviert [Reeder et al., 2007, Jöchl et al., 2008]. Die Sequenzen unterliegen dagegen Mutationen [Pevzner & Sze, 2000]. Solange die dreidimensionale Struktur beibehalten wird, ist die Sequenzebene zweitrangig. Dabei besteht in Helices eine kleinere Konservierung als in *Loop* Regionen, da diese häufig an Interaktionen beteiligt sind [Gräf et al., 2005]. Für das *Hammerhead* Ribozym bedeutet dies, dass eine korrekte Sequenz des Motivs nicht zwingend die typische Struktur des Motivs einnimmt, die für eine katalytische Aktivität benötigt wird.

Eine der größten RNA-Motiv-Datenbanken mit Kovarianz-Modellen und über 1400 RNA-Familien ist Rfam [Griffiths-Jones et al., 2003]. Neue mögliche Motive können mit bekannten Motiven verglichen werden und somit einen ersten Hinweis auf deren Funktionalität geben. Falls jedoch das Motiv nicht in der Datenbank enthalten oder die Ähnlichkeit zu gering ist, so dass es keine Übereinstimmung gibt, bleibt die experimentelle Validierung, bei der die Funktionalität der Struktur durch gezielte Mutation der Sequenz untersucht wird [Reeder et al., 2007].

1.5 Suchprogramme

Es gibt Suchprogramme, die wie in Abschnitt 1.4 beschrieben nach einem unbekannten Signal suchen und Programme, die ein gegebenes Motiv in einer Sequenz suchen. Zu den ersten zählen u. a. CONSENSUS [Hertz & Stormo, 1999], GibbsDNA [Lawrence et al., 1993], MEME [Bailey & Elkan, 1995], WINNOWER und SP-STAR, die in der Publikation [Pevzner & Sze, 2000] einander gegenübergestellt werden. Die meisten existierenden Suchalgorithmen, die auf Wahrscheinlichkeitsberechnungen und maschinellem Lernen basieren, konvergieren häufig nur gegen lokale Optima, die eher zufälligen *Pattern* entsprechen als tatsächlichen Signalen [Pevzner & Sze, 2000].

1.6 Wahrscheinlichkeiten

Die Beobachtung eines Motivs in einer zufälligen Sequenz entspricht einer bestimmten Wahrscheinlichkeit. Die Anzahl der zu erwartenden Treffer des Motivs kann mit dieser Wahrscheinlichkeit berechnet werden. Um die Wahrscheinlichkeit bestimmen zu können, müsste die Nukleotidhäufigkeit des jeweiligen Genoms berücksichtigt werden. Die folgenden Definitionen wurden aus [Gräf et al., 2005] übernommen. Zur Vereinfachung sei angenommen, dass die 4 Basen unabhängig voneinander sind und mit gleicher Wahrscheinlichkeit vorkommen. Die Wahrscheinlichkeiten pro Position werden miteinander multipliziert. Eine Sequenz der Länge 4, z. B. „ACTG“, besitzt somit eine Wahrscheinlichkeit von $p_{ACTG} = p_A * p_C * p_T * p_G = (1/4)^4 = 1/256 \approx 0.004$. Das bedeutet, dass alle 256 Nukleotide diese Sequenz 1 mal erwartet wird. Anders berechnet sich die Wahrscheinlichkeit einer Helix. Es gibt $2^4 = 16$ Möglichkeiten 2 Basen miteinander zu kombinieren, von denen 6 das Komplement zur Base bilden (4 Watson-Crick Basenpaare und 2 Wobble Basenpaare). Eine beliebige Helix der Länge 4 besitzt somit eine Wahrscheinlichkeit von $(6/16)^4 \approx 0.02$. Die Wahrscheinlichkeiten bei alternativen Strukturen im Fall variabler Helix- und *Loop*-Längen werden aufaddiert. Die Wahrscheinlichkeit eines RNA-Motivs bestehend aus Einzelsträngen und Helices errechnet sich aus dem Produkt der Einzelwahrscheinlichkeiten. Ein Beispiel hierzu befindet sich auf Seite 78.

1.7 Gibbs Energie

Die Umwandlung von Energien chemischer Reaktionen bei konstanter Temperatur und Druck werden mit Hilfe der Gibbs Energie G angegeben. Die Änderung der Energie vom Edukt zum Produkt wird als ΔG bezeichnet und setzt sich aus der Änderung der Enthalpie ΔH, der Temperatur T und der Änderung der Entropie ΔS zusammen ($\Delta G = \Delta H - T\Delta S$) [Markham & Zuker, 2005]. Ein negativer ΔG Wert entspricht einer Reaktion, die unter den gegebenen Bedingungen (Konzentrationen) freiwillig abläuft. Ist $\Delta G = 0$ oder $\Delta G > 0$ findet keine Reaktion statt, bzw. muss Energie hinzugefügt werden. Auf Strukturen übertragen ist der ΔG Wert ein Maß für deren thermodynamische Stabilität. Es ist die Energie, die notwendig ist, um eine Struktur zu entfalten [Tinoco et al., 2004].

1.8 ROC

Die *Receiver operating characteristic* (ROC) Kurve zeigt das Ergebnis einer binären Klassifizierung zur Auswahl eines optimalen Modells. Dazu wird die *True Positives* (TP) Rate (TP/(TP+ *False Negatives*)) gegen die *False Positives* (FP) Rate (FP/(FP+ *True Negatives*)) unter verschiedenen Einstellungen aufgetragen. Die beste Klassifizierung wird in der linken oberen Ecke bei einer Spezifität von 100% und einer Sensitivität von 100% erreicht. Eine zufällige Klassifizierung entspricht einem Punkt auf der Diagonalen [Sing et al., 2005].

1.9 Dateiformate

Die Daten liegen in verschiedenen Dateiformaten vor, welche an dieser Stelle kurz vorgestellt werden.

1.9.1 fa, fas, fasta, fastq

Sequenzinformationen und Sequenzen können im Fasta-Format gespeichert werden. Die Informationen, wie der Sequenzname, werden am Zeilenanfang mit einem “>” eingeleitet. In der darauffolgenden Zeile steht die Sequenz.

Das Programm RNAhit, welches in dieser Arbeit entwickelt wurde, nutzt u. a. eine abgewandelte Version des Fasta-Formates, um zusätzliche Informationen abspeichern zu können. Die Informationszeile beginnt ebenfalls mit dem “>” gefolgt von einem Sequenzbezeichner - meist der Sequenzregion, in der sie gefunden wurde. Alle weiteren Informationen, wie die Start- und Endposition, das verwendete Such- und Faltungsprogramm, die berechneten freien Energien, die dafür verwendeten *Constraints*, Nukleotidgehalt, Länge und so weiter (usw.) werden durch “:” voneinander getrennt. Ist die Startposition größer als die Endposition, so befindet sich die angegebene Sequenz auf dem Minusstrang. Der $\Delta\Delta G$ Wert setzt sich aus der Differenz der freien Energie der erzwungenen Struktur (ΔG_{motif}) und der freien Energie der frei gefalteten Struktur (ΔG_{free}) zusammen. In der nächsten Zeile folgt die Sequenz und in einer dritten Zeile die Sekundärstrukturinformation in Form einer Punkt-Klammer-Struktur.

Einen ähnlichen Aufbau, bestehend aus 3 bis 4 Zeilen pro Sequenz, besitzen fastq Dateien. Diese enthalten statt der Punkt-Klammer-Struktur den Phred Quality Score der einzelnen Nukleotide aus einer Sequenzierung [Cock et al., 2010].

1.9.2 gbk

GBK ist das GenBank Format, in dem neben der Sequenz zusätzliche Informationen und Annotationen gespeichert werden können. Ein Eintrag endet mit „//“ Zeichen. Weitere Informationen befinden sich unter
http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.

1.9.3 csv, tsv

Diese Formate ermöglichen einen tabellarischen Aufbau, bei dem die Spalten durch definierte Zeichen voneinander getrennt werden. CSV bedeutet *„comma-separated values“* und TSV - *„tab-separated values“*.

1.9.4 gff

Das *General Feature Format* (GFF) dient der allgemeinen Beschreibung von Eigenschaften, die mit DNA, RNA oder Proteinen assoziert werden. Jede Eigenschaft wird in einer Zeile definiert. Dazu zählen der Sequenzname, die Quelle, die Eigenschaft, die Start- und Endposition, der Score, der Strang und das Leseraster. Weitere Attribute und Kommentare sind optional. Nähere Informationen der Spezifizierung befinden sich unter (http://www.sanger.ac.uk/resources/software/gff/spec.html).

1.9.5 ini

Die ini Datei für den ProServer dient der Initialisierung wichtiger Referenzen. In dieser werden Pfade gesetzt und Voreinstellungen vorgenommen. Die Datei besteht aus mehreren Abschnitten mit Server spezifischen Optionen und Einstellungen der einzelnen *Distributed Annotation System* (DAS) Quellen (http://cpansearch.perl.org/src/RPETTETT/Bio-Das-ProServer-2.20/doc/proserver_guide.html). Eine Beispiel-ini-Datei des ProServers dieser Arbeit befindet sich im Anhang.

1.9.6 2bit

Das 2bit Dateiformat (http://genome.ucsc.edu/FAQ/FAQformat.html) ist eine Möglichkeit genomische Daten einfach und effizient zu komprimieren. Dafür werden die Nukleotide durch zwei Bits ersetzt (00 = T, 01 = C, 10 = A und 11 = G).

1.9.7 sam

Dieses flexible Format ermöglicht die Sicherung von *Alignment*-Informationen [Li et al., 2009]. Eine kurze Erklärung befindet sich im Bowtie2 Handbuch http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml oder unter http://samtools.sourceforge.net/. *Sequence Alignment Map* (SAM) Dateien besitzen Optional *Header*-Zeilen, die mit einem „@“ beginnen. Sie enthalten z. B. Informationen über die Referenzsequenz. Jede Zeile entspricht einem *Alignment* oder bei misslungenen *Alignments* einem *Read*. Eine Zeile besteht aus 12 Feldern, die durch Tabulatoren voneinander getrennt sind. Sie enthalten u. a. den Namen, eine Status ID, den Namen der Referenzsequenz, die *Mapping Quality*, die *Read*-Sequenz und weitere optionale Felder, wie *Alignment*-Informationen über die Anzahl *Mismatches, Gaps*, den Score und die Levenshtein-Distanz. Die Status ID 4 bedeutet z. B. nicht aligniert, 0 aligniert gegen den Plus- und 16 gegen den Minusstrang.

1.9.8 bam, bai

Die Abkürzung BAM steht für *Binary Alignment Map* und ist eine komprimierte Variante der SAM-Dateien [Li et al., 2009]. Zur Visualisierung dieser wird z. B. vom *Integrative Genomics Viewer* (IGV) eine sortierte Index Datei (bai) benötigt, die mit samtools erstellt werden kann.

1.10 Fragestellung

Die folgenden Fragen sollen im Rahmen dieser Arbeit beantwortet werden:

- In welchen Organismen kommen die bekannten Ribozym-Motive vor, wo sind sie lokalisiert und welche Variationen können gefunden werden?
- Wie können katalytisch aktive Ribozyme vorhergesagt werden?
- Welche Rückschlüsse lassen sich über die Evolution und mögliche Funktion dieser Motive ziehen?
- Bezüglich der *Deep Sequencing* Analyse stellte sich die Frage, ob RdRP, speziell RrpC, an der Produktion kleiner RNA und somit am RNAi-Mechanismus beteiligt ist und welchen Einfluss RdRP auf Retrotransposons hat.

In den nachfolgenden Kapiteln werden die verwendeten Materialien und Methoden vorgestellt, die zu den Ergebnissen in Kapitel 3 führten und anschließend diskutiert werden.

2 Material und Methoden

In diesem Kapitel werden die verwendeten Materialien (Hard-, Software) und Methoden vorgestellt.

2.1 Datenbankquellen

Die Sequenzdaten wurden aus verschiedenen öffentlichen Datenbanken heruntergeladen und lokal auf dem Server gespeichert. Dabei sollte beachtet werden, dass viele Datenbanken redundante Datensätze enthalten [Gräf et al., 2005] und es aufgrund unterschiedlicher Speziesnamen vorkommen kann, dass zwei gleiche Organismen aus verschiedenen Quellen heruntergeladen wurden, wie z. B. das Bakterium *Escherichia Coli*, dass bei NCBI als *„escherichia_coli“* und bei Ensemblgenomes als *„e_coli“* zu finden ist. Außerdem existieren mehrere Stämme, wie z. B. *„e_coli_bw2952“*, *„e_coli_bl21“* oder *„e_coli_dh10b“*, die sich ebenfalls namentlich unterscheiden. Die Gesamtdatenmenge ist in Abbildung 2.1 dargestellt und zeigt die geschätzte Größe des Suchraums.

Abbildung 2.1: Genomdaten
A) Gesamtanzahl aller Genome (Stand 08.09.2011). Diese Zahl steigt an, aufgrund der stetig wachsenden Anzahl neu sequenzierter Genome, die heruntergeladen und durchsucht werden können (siehe Abbildung 2.2). B) zeigt im Vergleich dazu das Größenverhältnis in Byte, welches bei vernachlässigter *Header*-Information der Fasta Dateien, in etwa der Größe des Suchraumes in Nukleotiden entspricht.

Die Daten wurden aus den folgenden Quellen heruntergeladen:

- ftp.sanger.ac.uk
- chgc.sh.cn
- ftp.ebi.ac.uk
- ftp.ensemblgenomes.org
- ftp.ensembl.org
- ftp.gramene.org
- ftp.jgi-psf.org
- ftp.ncbi.nlm.nih.gov
- jicbio.bbsrc.ac.uk
- medicago.org
- subviral.med.uottawa.ca
- hgdownload.cse.ucsc.edu
- genome.wustl.edu

Die vollständigen *Uniform Resource Locator* (URL)-Adressen der heruntergeladenen Genome befinden sich in der Datei „References.txt“ im Anhang. Die Genome wurden teilweise manuell und teilweise durch das in Abschnitt 3.1 vorgestellte Perl-Skript „getSpecies.pl“ heruntergeladen.

Die Anzahl sequenzierter Genome in den Datenbanken steigt stetig an (siehe Abbildung 2.2).

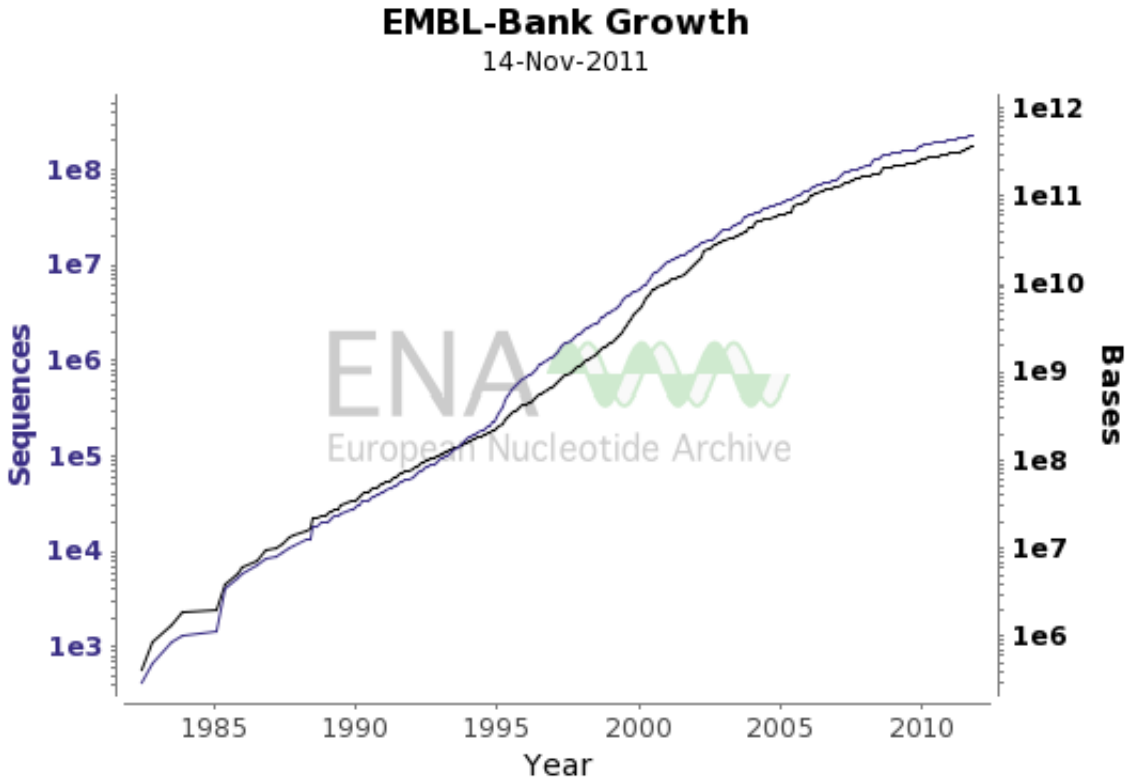


Abbildung 2.2: Embl Sequenzstatistik
Wachstum der Sequenzanzahl der Embl Datenbank in den letzten Jahrzehnten aus http://www.ebi.ac.uk/ena/about/statistics (Stand 20.11.2011).

Die Daten liegen in verschiedenen *Assembly*-Stufen vor, als *Contigs, Scaffolds* oder Chromosomen.

Die *Expressed Sequence Tag* (EST) Daten (dbEST) [Boguski et al., 1993] sind ein Teil von GenBank, die von National Center for Biotechnology Information (NCBI) (Stand 29.03.2010) heruntergeladen wurden. Sie enthalten Sequenzdaten von zahlreichen Organismen.

2.2 Zufallssequenzen

Da die Berechnung von Motivwahrscheinlichkeiten aufgrund variabler Helix- und *Loop*-Längen sehr komplex ist, wurden zur Abschätzung Zufallssequenzen in Größe der jeweiligen Genome generiert. Dabei sei zur Vereinfachung eine Gleichverteilung der Nukleotide angenommen, obwohl dies selten der Fall ist. Das Mittel der resultierenden Häufigkeiten des Motivs in den Zufallssequenzen dividiert durch die Sequenzlänge ergibt eine Approximation der Wahrscheinlichkeit. Werden genauso viel Treffer gefunden, wie statistisch erwartet, handelt es sich um Zufallstreffer. Die Sequenzen wurden mit Hilfe des Skriptes „gen_nucleic_stefan_graef.pl“ erzeugt (siehe Abschnitt 2.4.14).

2.3 Hardware

Als Entwicklungsumgebung dienten zwei verschiedene Systeme. Zum einen ein Notebook (Acer Aspire 5000 Series) mit folgenden Eigenschaften:

- AMD Turion™ 64 Mobile Technology ML-30 Prozessor 800 Mega Hertz (MHz),
- 960 Mega Byte (MB) *Random Access Memory* (RAM),
- 90,1 Giga Byte (GB) Festplatte

Das Betriebssystem ist Microsoft Windows XP Home Edition 5.1.2600, Service Pack 3 auf dem mit einem VMware® Player (Version 4.0.4) Ubuntu (Release 11.10 mit einem Linux 3.0.0 Kernel) simuliert wird. Zum anderen ein Server mit den folgenden Attributen:

- Asus *Dual-CPU Serverboard*, Modell DSBV-DX
- 2x Intel *Quadcore CPU* Xeon E5405 Prozessor, 2GHz
- Arbeitsspeicher 4 GB *Kit* (2 x 2 GB) Kingston
- 2x 500 GB Festplatte Seagate Barracuda 7200.12, Modell ST3500410AS

Das Betriebssystem ist Ubuntu (9.04 mit einem Linux 2.6.27-11-Server Kernel). Auf diesem System wurden alle Suchen ausgeführt.

Um sämtliche Funktionen des im Rahmen der Arbeit entwickelten Programmes RNAhit nutzen zu können, müssen folgende Programme, die zum Teil im nächsten Abschnitt näher erläutert werden, installiert sein. Es wird mindestens ein Suchprogramm PatScan 1.01 [Dsouza et al., 1997] mit squid-1.9g [Eddy, 2005b] oder RNAbob [Eddy, 2005a] und ein Faltungsprogramm Mfold [Zuker, 2003] oder UNAfold [Markham & Zuker, 2008] benötigt.

Außerdem sollte die folgende Umgebungsvariable, je nach verwendeter Shell und Betriebssystem, in diesem Fall in der .bashrc Datei gesetzt werden.

```
PERL5LIB=/media/sdb2009/src/bioperl-live:
/media/sdb2009/src/ensembl/modules:
/media/sdb2009/src/ensembl-compara/modules:
/media/sdb2009/src/ensembl-functgenomics/modules:
/media/sdb2009/src/ensembl-variation/modules:
/root/.cpan/build:
/usr/lib/perl/5.10.0:
/usr/lib/perl5
export PERL5LIB
```

Die Ensembl Schnittstelle verlangt weitere zu setzende Umgebungsvariablen. Diese befinden sich in der Datei efg im Anhang.

Da die Such- und Faltungsprogramme auf Linux basieren, ist Linux das Betriebssystem der Wahl.

Des Weiteren sollte optional eine MySQL-Datenbank mit installiertem Ensembl-Schema vorhanden sein, falls Daten in dieser gespeichert werden sollen. Auf dem Server ist Ensembl Release 53 installiert. Inzwischen existiert Release 68 (Stand 17.10.2012). Das Datenbankschema kann über MySQL in eine neu erstellte Datenbank geladen werden. Die Datei table.sql befindet sich zusammen mit verschiedenen Patches im „ensembl/sql/“ Ordner und im Anhang.

```
mysql -u root -p mysql
CREATE DATABASE HyPaDB_core;
mysql -u root -p HyPaDB_core < table.sql
```

Das Speichern der Ergebnisse in der Datenbank geschieht mit Hilfe der EnsemblAPI (Abschnitt 2.4.15).

Die Auswertung der *Deep Sequencing* Daten erfolgte über die Swedish National Infrastructure for Computing (SNIC) im Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) im Rahmen des Projektes b2011213 in einem Cluster (http://www.uppmax.uu.se/systems/kalkyl) bestehend aus

- 348 HP SL170h G6 Servern
  - jeder Server besteht aus 2 Quadcore Intel Xeon 5520 (Nehalem 2.26 Giga Hertz (GHz), 8 Mega Byte (MB) Cache) Prozessoren
- 348 Knoten mit 696 Prozessoren *à* 4 Kernen = 2784 64-bit Prozessorkerne
  - 316 Knoten *à* 24 Giga Byte (GB) Speicher und 250 GB Festplatte
  - 16 Knoten *à* 48 GB Speicher und 250 GB Festplatte
  - 16 Knoten *à* 72 GB Speicher und 2 Tera Byte (TB) Festplatte
- 9504 Giga Byte (GB) RAM
- 113 Tera Byte (TB) Festplatte
- miteinander verbunden über ein 4:1 DDR InfiniBand fabric,

in Zusammenarbeit mit Johan Reimegård und Fredrik Söderbom.

2.4 Software

2.4.1 gedit

Dieses Programm in Version 2.30.3 (http://projects.gnome.org/gedit/) ist ein Linux Texteditor, mit dem die unterschiedlichsten Skripte erstellt werden können.

2.4.2 Gimp

Gimp (http://www.gimp.org/) steht für „GNU Image Manipulation Program“ und ist ein Bildbearbeitungsprogramm. Es liegt in Version 2.6.6 vor und dient der Erstellung und Bearbeitung von Grafiken.

2.4.3 KolourPaint

KolourPaint (http://www.kolourpaint.org/) Version 4.7.4 ist ein Linux Mal- und Zeichenprogramm.

2.4.4 Paint

Ein weiteres Mal- und Zeichenprogramm zum Anzeigen, Erstellen und Bearbeiten von Bildern ist Microsoft®Paint (Version 5.1.).

2.4.5 Jmol

Jmol (http://jmol.sourceforge.net/) in Version 12.0.41 ist ein *Open Source* Molekül Betrachter zur dreidimensionalen Visualisierung von chemischen Strukturen. Jmol wurde verwendet, um die Kristallstrukturen darzustellen und die Abbildungen 1.1, 1.6(a) und 1.6(b) zu erzeugen.

2.4.6 BioEdit

BioEdit in Version 7.0.5 ist ein Windows Sequenz-*Alignment*-Editor und Analyseprogramm [Hall, 1999] und wurde für die manuelle Korrektur des *Multiple Sequence Alignments* (MSA) (Abbildung 3.19) verwendet.

2.4.7 CLC sequence viewer

Der CLC Sequence Viewer (http://www.clcbio.com/index.php) in Version 6.2.64 ermöglicht zahlreiche Bioinformatik-Analysen kombiniert mit Datenmanagement und verschiedenen Ausgabemöglichkeiten.
Darunter zählen z. B. MSA-Algorithmen, wie ClustalW [Thompson et al., 1994], MUSCLE [Edgar, 2004] und Kalign [Lassmann & Sonnhammer, 2005] sowie Phylogenie Programme, wie z. B. Neighbor Joining [Saitou & Nei, 1987] und UPGMA [Sokal & Michener, 1958]. Das Programm diente dem Vergleich verschiedener MSA und der Erstellung eines phylogenetischen Baums (Abbildung 3.20) mit Hilfe des Neighbor Joining Algorithmus und 1000 Bootstrap Wiederholungen.

ClustalW

ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/) ist ein MSA-Programm, das mehrere Sequenzen untereinander aligniert, um konservierte Regionen zu ermitteln. Dazu werden zwischen allen Sequenzen paarweise *Alignment* Scores berechnet, die sich aus der Anzahl der *Matches* dividiert durch die Länge des *Alignments* (ohne *Gaps*) ergeben. Diese Scores (prozentuale Identität) werden in Distanzen umgerechnet (1-(Score/100)). Daraus kann mit dem Neighbor Joining Algorithmus [Saitou & Nei, 1987] ein Ähnlichkeitsbaum erzeugt werden, in dem minimale Abstände zwischen Distanzen in Gruppen zusammengefasst werden. Dieser Baum dient im dritten Schritt der Reihenfolge, in der die *Alignments* zusammengefügt werden [Thompson et al., 1994].

Neighbor Joining

Der Neighbor Joining Algorithmus [Saitou & Nei, 1987] modifiziert von Studier und Keppler [Studier & Keppler, 1988] erstellt aus einer Distanzmatrix einen ungewurzelten phylogenetischen Baum mit entsprechenden Kantenlängen. Zu Beginn hat der Baum eine sternförmige Struktur. In diesem werden anschließend Paare (Nachbarn, Blätter) zusammengefasst, die einen minimalen Abstand zueinander besitzen. Ihr gemeinsamer Vorfahr wird als neuer Knoten mit neuen Kanten hinzugefügt und die alten Kanten entfernt. Nach jedem Schritt wird die angepasste Distanzmatrix mit Hilfe des mittleren Abstands aller Knoten zum neuen Knoten neu berechnet. Dies wird solange wiederholt bis zwei Knoten verbleiben [Saitou & Nei, 1987, Studier & Keppler, 1988].

Bootstrapping

Beim Bootstrapping wird ein neues MSA aus den Spalten des Original MSA mit zurücklegen erzeugt und anschließend ein neuer Baum erstellt. Die Bootstrap-Werte an den Kanten entsprechen dem relativen Anteil an Bäumen, welche die gleiche Kante wie der Originalbaum besaß [Zharkikh & Li, 1992].

2.4.8 PHPmyAdmin

PHPmyAdmin (http://www.phpmyadmin.net/home_page/index.php) Version 3.1.3 ermöglicht die Administration einer MySQL-Datenbank über eine PHP: ***H**ypertext **P**reprocessor* (PHP) Webseite. Abbildung 2.3 zeigt beispielhaft die Rechteverwaltung der lokalen Datenbank.

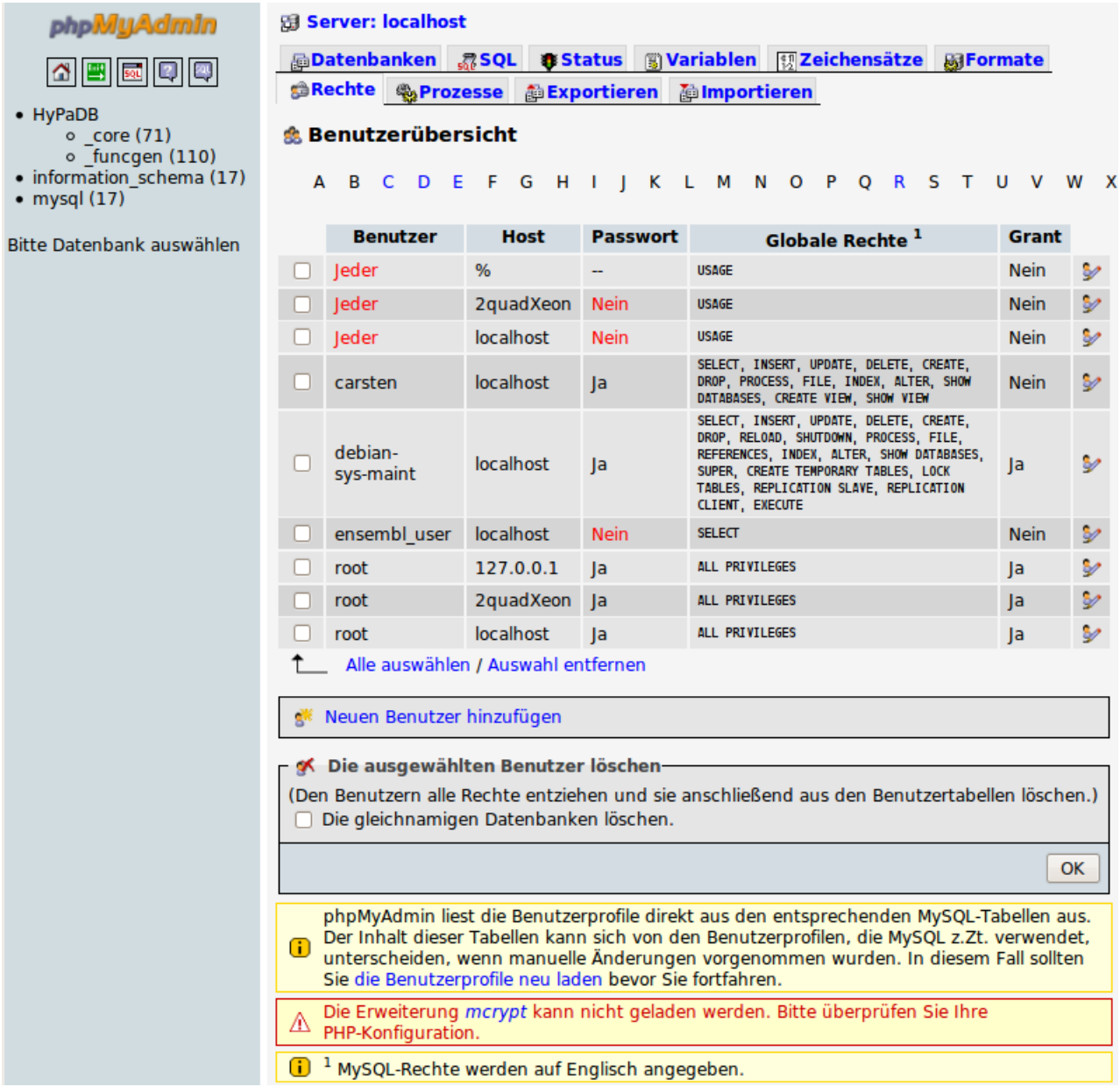


Abbildung 2.3: PHPmyAdmin
Darstellung der Rechteverwaltung der Benutzer.

2.4.9 Perl

Perl (http://www.perl.org/) liegt in Version 5.10.1 vor. Es eignet sich als Programmiersprache zur Bearbeitung der Fragestellungen, da z. B. die Schnittstelle mit Ensembl ebenfalls auf Perl basiert.

Perlbibliotheken

BioPerl ist eine umfangreiche Bibliothek mit vielfältigen Funktionen. Die EnsemblAPI benötigt BioPerl in Version 1.2.3, da spätere Versionen inkompatibel sind. Des Weiteren wird eine Datenbankschnittstelle zwischen Perl und MySQL benötigt, die in dem Perl-Modul DBI (http://dbi.perl.org/) enthalten ist. Für die Erstellung der grafischen Oberfläche wird GTK2 (http://gtk2-perl.sourceforge.net/) eingebunden.

CPAN

Fehlende Perl-Module können vom „Comprehensive Perl Archive Network“ (CPAN) (http://search.cpan.org/) nach installiert werden.

2.4.10 R

R (http://www.r-project.org/) in Version 2.8.1 [Team, 2008] ist eine Skriptsprache und Umgebung zur Erstellung von Statistiken und Grafiken. Es bietet zahlreiche Analysetechniken und ist durch optionale Bibliotheken jederzeit erweiterbar. Diese können vom „Comprehensive R Archive Network“ (CRAN) (http://cran.r-project.org/) heruntergeladen werden.

2.4.11 MySQL

Auf dem Server ist die MySQL-Server Version 14.12 Distribution 5.0.75 (http://www.mysql.de/) installiert, um die Suchergebnisse speichern zu können. Auf dem Notebook ist lediglich ein MySQL-Client Version 14.14 Distribution 5.1.54 installiert, um eine Verbindung zur MySQL-Datenbank herstellen zu können. Mit dem Befehl mysqldump kann eine Sicherung der Datenbank erstellt werden. Dazu müssen zunächst alle Umgebungsvariablen geladen werden. Anschließend kann mit dem Befehl:

```
mysqldump $MYSQL_ARGS -uroot -p$1 --add-drop-table $COLL_DBNAME
      | gzip -c > /media/sda2009/src/backup/$COLL_DBNAME.$(date -I).sql.gz
```

ein Backup angelegt werden. Falls ein Backup in eine bestehende Datenbank geladen werden soll, muss der Inhalt zunächst gelöscht werden.

```
mysql -h localhost -u carsten -p HyPaDB_core < deleteDB.sql
```

Da die „meta“ Tabelle verschiedene Patches enthält, müssen die Spezies manuell entfernt werden. Dazu eignet sich die Bearbeitung mit PHPmyAdmin. Das Backup kann anschließend mit

```
mysql -h localhost -uroot -p HyPaDB_core < HyPaDB_core.2011-01-07.sql
```

geladen werden.

2.4.12 Suchprogramme

Es gibt zwei Arten nach einem gegebenem Motiv zu suchen. Zum einen wird direkt in der Sequenz oder in zuvor berechneten Indizes nach dem Motiv gesucht, wobei Index basierte Programme gewöhnlich schneller sind [Gräf et al., 2005]. Zum anderen gibt es Suchprogramme, die das Such-*Pattern* analysieren, um statistisch signifikante Teile des *Patterns* zu bestimmen und diese effizienter suchen zu können [Gräf et al., 2005].

Nachfolgend werden mehrere getestete Suchprogramme vorgestellt.

PatScan

Das erste Suchprogramm, welches eine Sequenz- und Sekundärstruktursuche ermöglicht, wurde von Forschern des Argonne National Laboratory entwickelt und heißt PatScan [Dsouza et al., 1997]. Eine von Pesole *et al*. überarbeitete Version heißt PatSearch [Pesole et al., 2000, Grillo et al., 2003].

Bei der hier verwendeten Version von PatScan handelt es sich um eine von Stephan Zanger im Rahmen seiner Diplomarbeit modifizierte Variante [Zanger, 2005]. Diese ermöglicht zusätzlich:

- das Einlesen zahlreicher Dateiformate mit ZLib (http://zlib.net/) ohne vorherige Konvertierung,
- eine für spätere Zwecke sehr nützliche Ausgabe der Punkt-Klammer-Struktur eines Treffers,
- eine automatische Speicherverwaltung
- sowie die Ausgabe überlappender Treffer [Zanger, 2005].

PatScan verwendet für die Suche ein in einer Datei beschriebenes Muster, um in einer weiteren Datei, die ein oder mehrere Sequenzen enthält, nach dem Motiv zu suchen. Es nutzt eine modulare Sprache mit der Helices und Einzelstränge beliebiger Länge definiert werden können. Des Weiteren gibt es verschiedene Optionen, welche die Suche und Ausgabe der Ergebnisse beeinflussen. So kann z. B. auf dem komplementären Strang mit (-c) oder überlappend mit (-o) gesucht werden und eine Punkt-Klammer-Notation (-s), Leerzeichen zwischen den Musterelementen und zusätzliche Nukleotide auf beiden Seiten des Motivs ausgegeben werden. Es ist auch möglich nach einer bestimmten Anzahl von Treffern (-m) oder Fehlschlägen (-n) die Suche abzubrechen oder ausgewählte Sequenzen mit (-i) zu überspringen. Der Aufruf und sämtliche Optionen können mit

```
./patscan -h
```

abgefragt werden. Ein Beispiel für eine Musterbeschreibungsdatei befindet sich in Abbildung 2.7(a). Sie besteht aus:

- Kommentarzeilen, die mit einem „%“ Zeichen beginnen, z. B. % ID hammerhead
  - als Erweiterung im Rahmen dieser Arbeit besteht die Möglichkeit eigene Definitionen für anschließende Analysen (Faltungen) mit „% PH“ einzubauen, z. B. % PH s3=CUGANGA s5=GAA,
- Paarungsregeln, die festlegen, nach welchen Kriterien komplementäre Basen Basenpaarungen eingehen dürfen, z. B. r1={au,ua,gc,cg,gu,ug} für Watson-Crick und Wobble Basenpaarungen,
- Einzelstrangregionen, die mit einer von / bis Angabe oder durch IUPAC-Symbole [Cornish-Bowden, 1985] (Tabelle 1.1) definiert werden können, z. B. UH, wobei auch Fehlerangaben möglich sind (*Mismatches*, Deletion, Insertion), z. B. UH[1,0,1], was ebenfalls für Helices gilt und
- Helices, welche mit „px“ und einer von / bis Länge definiert werden. Komplementäre Basen werden mit Hilfe der Paarungsregeln gefunden, z. B. p1=3...6 4...5 r1~p1 definiert eine Helix der Länge 3 bis 6 mit einem *Loop* aus 4 bis 5 beliebigen Nukleotiden.

Wenn alle Kriterien erfüllt sind, wird ein Treffer wie folgt, am Beispiel eines Typ III *Hammerhead* Ribozyms, ausgegeben:

```
>SJC_S000033:[571280,571235]
TTGTATCTTCTTAAAGATTGATAATGGTTGACATAACTGGAAACAA
(((..((((....))))(....).((((.......))))....)))
```

Zu sehen ist ein Fasta ähnliches Format (siehe Abschnitt 1.9.1).

RNAbob

Dieses Suchprogramm (ftp://selab.janelia.org/pub/software/rnabob/), welches von Sean Eddy entwickelt wurde, ist eine modifizierte Variante von RNAMOT [Laferriere et al., 1994] mit einem anderen zugrunde liegendem Algorithmus. Es verwendet für die Suche nach Sequenz- und Sekundärstrukturübereinstimmungen einen graphentheoretischen Ansatz [Eddy, 2005a].

Der Aufbau der Musterbeschreibungsdatei [Webb et al., 2009] besteht aus einer Anordnung verschiedener Musterelemente, mit einer anschließenden Definition der einzelnen Elemente. Helices werden mit h oder r gekennzeichnet, wobei h Watson-Crick und Wobble Basenpaare erlaubt und r ausschließlich Watson-Crick Basenpaare zulässt. Die komplementäre Base wird mit h’ bzw. r’ definiert. Einzelstrangregionen werden mit s beschrieben [Webb et al., 2009].

Die Definition der Basen erfolgt durch IUPAC Symbole [Cornish-Bowden, 1985] (Tabelle 1.1). Kommentare besitzen am Zeilenanfang ein #-Zeichen. Zusätzlich können für spätere Faltungen mit # PH eigene Faltungsbeschränkungen definiert werden. Bei der Definition der Helices und der Einzelelemente können *Mismatches* erlaubt werden. Alternative Sequenzpositionen werden durch * oder der Anzahl in eckigen Klammern symbolisiert. Eine Beispieldefinition befindet sich in Abbildung 2.7(b). Ergebnisse werden in der folgenden Form ausgegeben (gleiches Beispiel aus dem PatScan Abschnitt):

```
571280 571235  SJC_S000033
|TTG|TA|TCTT|CTTA|AAGA|T|TGAT|A|A|TGGT|TGACATA|ACTG|GAAA|CAA|
```

Zu sehen sind die Start-, Endposition und die Region sowie eine Zeile darunter die Sequenz des Treffers, die entsprechend dem Such-*Pattern* durch „|“ voneinander getrennt ist.

Vmatch

Vmatch (http://www.vmatch.de/) ist eine Sequenzanalyse Software von Stefan Kurtz, die effizient *Mapping*-Aufgaben erfüllt. Dazu wird aus der Referenzsequenz ein Index (*Suffix Array* [Manber & Myers, 1993]) erzeugt, der konstante Bereiche zusammenfasst. Anschließend kann ein *Pattern* gesucht und die Position des *Pattern* ermittelt werden.

blastall

Die *Basic Local Alignment Search Tool* (BLAST) Programme (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/) sind häufig verwendete Werkzeuge zur Suche ähnlicher Sequenzen in Datenbanken. BLAST ist eine Heuristik, die zwischen gegebener Sequenz und Datenbank ein lokales *Alignment* erstellt. Es besteht aus zwei Teilen, einem Suchalgorithmus und der Berechnung der statistischen Signifikanz. Dazu werden lokal optimale Paare (HSPs) ermittelt, welche solange in 5’ und 3’ Richtung erweitert werden, bis der Score sich nicht mehr verbessert. Alle Treffer, die größer als ein bestimmter Schwellwert sind, werden ausgegeben [Altschul et al., 1997]. Für eine BLAST Suche muss zuvor mit formatdb ein Index angelegt werden. BLASTN ist die Homologiesuche nach Nukleotidsequenzen. Ensembl verwendet für die Abbildung von Sequenzen zwischen zwei Genomen BLASTZ [Schwartz et al., 2003]. Dafür werden zunächst *Repeats* entfernt und anschließend identische 12-mere gesucht, die erweitert werden und ab einem bestimmten Score *Gaps* erlauben. In einem zweiten Schritt werden die gefundenen Paare mit sensitiveren Parametern durchsucht, in dem z. B. nach identischen 7-meren gesucht und die Schwelle des Scores zum Erlauben von *Gaps* herabgesetzt wird.

seedtop

Ein weiteres Programm aus dem BLAST Paket ist seedtop von Tao Tao
(http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/seedtop.html).
Es ist ein *Pattern* Suchprogramm, welches ähnlich zu BLAST einen mit formatdb angelegten Index benötigt. Ein Beispielaufruf von seedtop lautet:

```
seedtop -p pattern -f -d chr4.fa -k test.txt -o test.out
```

Das *Pattern* für ein *Hammerhead* Ribozym kann wie folgt definiert werden:

```
[ACGT](3,6)-T-{G}-x(12,20)-C-T-G-A-[ACGT]-G-A-x(12,20)-G-A-A-A-[ACGT](3,6).
```

Eckige Klammern beziehen sich auf eine einzelne Position, die einem der enthaltenen Nukleotide entspricht. Runde Klammern markieren einen Bereich mit von / bis Angaben. Einzelne Positionen werden durch einen Bindestrich voneinander getrennt. Geschweifte Klammern beinhalten für die Position auszuschließende Nukleotide. Ein Punkt markiert das Ende des *Patterns*. Die Ausgabe eines Treffers erfolgt über die Position des *Patterns*.

SQUID

SQUID Version 1.9g (http://selab.janelia.org/software.html) [Eddy, 2005b] ist eine C-Bibliothek von Sean Eddy, die Funktionen zur Sequenzanalyse anbietet. SQUID wird von PatScan benötigt.

miR-abela

Das Programm miR-abela ist ein Vorhersageprogramm für miRNA und sucht nach möglichen *precursor* miRNA [Sewer et al., 2005].

2.4.13 Faltungsprogramme

Für die Sekundärstrukturvorhersage und zur Berechnung der minimalen freien Energie wurden verschiedene Faltungsprogramme getestet.

PseudoViewer

Dieses Programm (http://wilab.inha.ac.kr/pseudoviewer/) von Byun und Han in Version 3.0 [Byun & Han, 2009] ermöglicht die Visualisierung von gegebenen RNA-Sekundärstrukturen und Pseudoknoten.

pknotsRG

Das Programm pknotsRG Version 1.3 (http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/) ist ein RNA-Faltungsprogramm, welches zusätzlich einfache Pseudoknoten berücksichtigen kann. Die Energieparameter für Strukturen, die keine Pseudoknoten beinhalten, sind die gleichen wie bei Mfold 3.1 [Reeder & Giegerich, 2004].

kinefold

Mit kinefold (http://kinefold.curie.fr/) können ebenfalls RNA-Faltungsvorhersagen generiert werden, die Pseudoknoten berücksichtigen. Es bietet zwei Faltungsoptionen. Die erste erlaubt eine renaturierende Faltung, beginnend aus einem denaturierten Zustand und die zweite ermöglicht eine co-transkriptionale Faltung, bei der nach und nach die Nukleotide hinzugefügt werden. Des Weiteren besteht die Möglichkeit definierte Einzelstränge und Helices zu erlauben bzw. zu verbieten [Xayaphoummine et al., 2005].

Mfold

Mfold in Version 3.5 ist ein Nukleinsäure Faltungs- und Hybridisierungsprogramm [Zuker, 2003], welches die Energie-Parametereinstellungen aus [Mathews et al., 1999] verwendet. Mfold besitzt folgende Standardeinstellungen (http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form):

- die RNA-Sequenz ist linear,
- 37°C Faltungstemperatur,
- 1 Molar $[Na^+]$, 0 Molar $[Mg^{2+}]$,
- 5% suboptimale Strukturen werden berechnet,
- insgesamt < 50 Faltungen,
- Standard Fensterparameter entsprechend der Sequenzlänge,
- < 30 nt *internal / bulge Loop*-Größe und
- ohne maximalen Abstand zwischen Basenpaaren.

Es erfolgt eine detaillierte Ausgabe der Struktur, Einzelstranghäufigkeit und Energie [Zuker, 2003]. Mfold ermöglicht neben der Berechnung der minimalen freien Energie, die Berechnung der freien Energie einer erzwungenen Struktur. Dazu werden beim Aufruf *Constraints* übergeben, die Basenpaarungen von Positionen erzwingen (F) oder verhindern (P).

Vier Parameter sind für eine Bedingung notwendig:

- „F“ oder „P“ für das Erzwingen oder Verbieten,
- die Startposition,
- die Endposition
- und die Anzahl einer oder mehrerer Wiederholungen.

Zum Beispiel lauten die *Constraints* zur folgenden HHRz III Punkt-Klammer-Struktur wie folgt:

```
ACCGGTCTGTAGTTGGGAATCTTGACTTGTTTGTGAATCCCCACTCCTGGCTCTACCT
(((((..((((.............................................
AAGCTGATCAGAACACCCTGCACTACACTGATGAGCCCCCAAGAAGGGCGAAACCGGT
.......................))))(....).((((.......))))....)))))
F 1 116 5
F 8 85 4
P 86 0 7
F 93 107 4
P 108 0 3
```

Hierbei sollen das Nukleotid A auf Position 1 mit dem Nukleotid T auf Position 116, Nukleotid C auf Position 2 mit Nukleotid G auf Position 115 und die weiteren 3 Nukleotide Basenpaarungen eingehen. Das gleiche für die Positionen 8 bis 12 mit den Positionen 81 bis 85. Helix II des HHRz wird nach dem gleichen Prinzip erzwungen. Die Position 86 und die darauf folgenden 7 Positionen sollen keine Basenpaarungen bei der Faltung eingehen. Das gleiche gilt für die Positionen 108 bis 111.

Die automatische Generierung der *Constraints* in dem Programm RNAhit geschieht mit Hilfe der Punkt-Klammer-Struktur. Dazu werden die Sequenz und die Punkt-Klammer-Struktur in einzelne Zeichen (Character) zerlegt. Anschließend werden die Positionen der Character in einen Stapel (Stack) geschoben, wobei die erste schließende Klammer mit der letzen öffnenden Klammer erzwungen wird. Auf Sequenzebene entsprechen diese Positionen einem Basenpaar. Die Positionen der Punkte werden ebenfalls gespeichert. Aufeinanderfolgende Positionen werden danach gezählt und durch den vierten Parameter der Bedingung zusammengefasst.

Das Erzwingen von Helices bestehend aus einem oder zwei Basenpaaren, wie im obigen Beispiel die potenzielle Bedingung F 86 91 1, ist in Mfold nicht möglich. Um dennoch die erzwungene freie Energie dieser Struktur automatisch berechnen zu können, gibt es die Möglichkeit in der Musterbeschreibungsdatei des Suchprogrammes *Constraints* für RNAhit selbst zu definieren, wodurch die Komplexität der Generierung steigt. Die oben erwähnten Positionen werden in diesem Fall in einem zweiten Schritt überarbeitet. Weitere ausführliche Informationen befinden sich im RNAhit Quellcode. Die verschiedenen Faltungseinschränkungen sind in Tabelle 2.2 dargestellt. Eine Einschränkung wie „h3=N s3=UGAN h3'=N s4=A s6=GAA“ für das HHRz Typ III bedeutet, dass Helix 3 aus der RNAbob-Beschreibung nicht Erzwungen werden darf und diese Bedingung entfernt wird. Die Einzelstränge s3, s4, s6 sollen ungepaart bleiben, wobei im Unterschied zur Motivbeschreibung in s6 lediglich „GAA“ ungepaart bleiben soll und A15.1 mit U16.1 eine Basenpaarung eingehen darf.

UNAfold

UNAfold in Version 3.6.1 (http://mfold.rna.albany.edu/?q=DINAMelt/software) steht für „Unified Nucleic Acid Folding“ und ist die Weiterentwicklung von Mfold. UNAfold ist ein Software-Paket mit verschiedenen Programmen zur Faltungs-, Hybridisierungs- und Schmelzsimulation von einer oder zwei Nukleinsäuresequenzen [Markham & Zuker, 2008]. Speziell das Programm hybrid-ss-min ermöglicht die Berechnung der minimalen Faltungsenergie einer Sequenz. Diese wird für steigende Temperaturen zwischen minimaler und maximaler Temperatur berechnet bzw. für eine einzelne Temperatur, wenn die minimale und maximale Temperatur gleich hoch ist. Die freie Energie wird in die Datei „*.dG“ geschrieben und kann dort ausgelesen werden. Die Option -c ermöglicht die Angabe von eigenen *Constraints* im gleichen Format wie bei Mfold, die automatisch der „*.aux Datei“ entnommen werden.

UNAfold besitzt die folgenden Standardeinstellungen:

- eine RNA-Sequenz ist gegeben,
- die minimale Temperatur beträgt 37°C,
- die Temperatur steigt in 1°C Schritten,
- die maximale Temperatur beträgt 37°C,
- 1 Molar $[Na^+]$, 0 Molar $[Mg^{2+}]$,
- verwende Salzkorrektur für Polymere statt für Oligomere und
- $<$ 30 nt *internal / bulge Loop*-Größe [Markham & Zuker, 2008].

RNAfold

RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) ist ein weiteres Faltungsprogramm, das für eine gegebene Sequenz auf Grundlage der minimierten freien Energie eine Sekundärstruktur vorhersagen kann [Hofacker et al., 1994]. Es bietet Struktur *Constraints*, verschiedene Faltungs- und Energieparameter sowie mehrere Ausgabemöglichkeiten.

2.4.14 Skripte

Es wurden die folgenden C++ und Perl-Skripte verwendet.

C++

twoBitToFa

Das C++ Skript twoBitToFa (http://genome.ucsc.edu/FAQ/FAQdownloads.html) wurde von der University of California Santa Cruz (UCSC) entwickelt und wandelt 2bit Dateien in Fasta Dateien um. Es wurde verwendet, um die im 2bit Format heruntergeladenen Genome des UCSC durchsuchen zu können.

Perl

gen_nuclic_stefan_graef

Dieses Perl-Skript aus dem HyPa-Paket [Gräf, 2005] diente der Erstellung von Zufallssequenzen. Die Erzeugung von DNA-Sequenzen mit einer gleichverteilten Nukleotidhäufigkeit ist voreingestellt. Optional kann eine gegebene Anzahl von RNA-Sequenzen mit einer definierten Mindest- und Maximallänge sowie modifizierten Nukleotidhäufigkeit generiert werden.

Aufgrund der Größe des Arbeitsspeichers war eine Generierung von mehr als 2 Milliarden Nukleotiden nicht möglich.

2.4.15 Ensembl API

Die Ensembl *Application Programming Interface* (API) (http://www.ensembl.org/info/docs/api/index.html) ist eine Perl-basierte Schnittstelle zwischen den Datenbanken und spezifischen Programmen. Sie ermöglicht einen effizienten Zugriff auf die Daten, um Sequenzen automatisch zu annotieren, untereinander zu vergleichen und Variationen sowie Regulationen zu bestimmen. Eine Installationsanleitung wird auf der Ensembl-Seite unter ../api/api_installation.html genau beschrieben. Die Ensembl API liegt in Release 53 vor. Es ist wichtig, dass sowohl die API als auch die Datenbank von der gleichen Release sind.

2.4.16 ProServer

Der ProServer (http://www.sanger.ac.uk/resources/software/proserver/) [Finn et al., 2007] ist ein in Perl programmierter DAS-Server. Dieser erlaubt über das DAS-Protokoll den Austausch von Daten zwischen mehreren Datenbanken. Die ProServer Architektur ist in Abbildung 2.4 zu sehen.

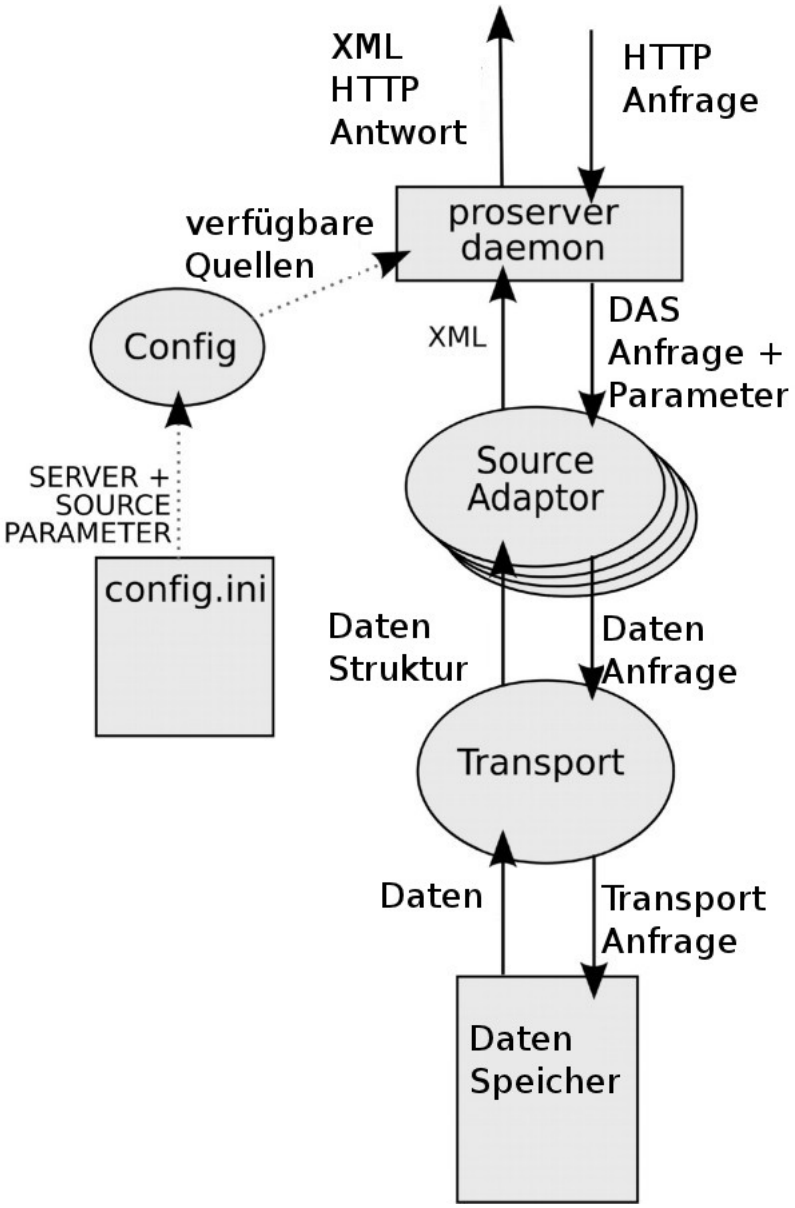


Abbildung 2.4: ProServer Architektur
2.4 zeigt den Aufbau und die Funktionsweise eines ProServers, angepasst aus [Finn et al., 2007].

Mit Hilfe einer Konfigurationsdatei können mehrere Quellen (z. B. das Dateisystem und / oder eine Datenbank) definiert werden. Eine über Hypertext Transfer Protokoll (HTTP) ankommende Anfrage an den ProServer Daemon wird in eine DAS Anfrage für den Source Adaptor übersetzt, welcher daraufhin über den Transport Adaptor eine *Structured Query Language* (SQL) Abfrage an die Datenbank sendet und die erhaltenen Daten in eine *Extensible Markup Language* (XML)-Datei umwandelt. Diese wird anschließend mit Hilfe von *Extensible Stylesheet Language* (XSL)-*StyleSheets* auf dem Bildschirm ausgegeben [Finn et al., 2007]. Der Start erfolgt mit dem Befehl:

```
perl eg/proserver -x -c eg/hypaserver.ini
```

2.4.17 Suchen

Der erste Schritt in der Motivsuche war die Motivbeschreibung. Anschließend wurden zahlreiche Motivsuchen durchgeführt. Die einzelnen Definitionen und Parametereinstellungen werden in den folgenden zwei Abschnitten vorgestellt.

Motivdesign

Um ein Motiv definieren zu können, werden bekannte Sequenzen benötigt, in denen das Motiv enthalten ist. Falls nicht genügend Beispiele vorhanden sind, kann nach ähnlichen Sequenzen durch eine Homologiesuche, z. B. mit BLAST [Altschul et al., 1997] oder FastA [Pearson, 1990], gesucht werden. Dabei sollte beachtet werden, dass möglichst keine falsch positiven Sequenzen (FP) hinzugefügt werden, da dies zu fehlerhaften Schlussfolgerungen im Motiv führen kann. Aus den Sequenzen wird anschließend ein MSA erzeugt, in dem strukturelle Informationen beim Anordnen der Sequenzen berücksichtigt werden sollten. Dies ist z. B. mit MARNA [Siebert & Backofen, 2005] oder limitiert mit RE-MuSiC [Chung et al., 2007] möglich. Aus dem MSA kann anschließend eine Konsensussequenz bestimmt werden. Die Konsensussequenz fasst mit Hilfe der IUPAC Symbole [Cornish-Bowden, 1985] die unterschiedlichen Nukleotide pro Position zusammen. Die daraus entstandene allgemeine Beschreibung des MSA entspricht dem Motiv. Eine genomweite Suche des Motivs sollte mindestens alle zur Definition verwendeten Sequenzen finden. Dazu eignet sich am besten ein Chromosom oder Genom, das bereits vollständig sequenziert und annotiert ist, da auf diese Weise zusätzlich gefundene Treffer besser bestimmt werden können. In den darauffolgenden Schritten sollten weitere Informationen in die Bewertung der gefundenen Treffer einfließen, bevor die Motive experimentell verifiziert werden [Gräf et al., 2005].

Das folgende Beispiel soll diese Vorgehensweise demonstrieren. Ziel ist, einen HHRz I Deskriptor zu erzeugen. Es gibt bekannte Sequenzen aus den Veröffentlichungen von Rojas [Rojas et al., 2000], Ferbeyre [Ferbeyre et al., 1998, Ferbeyre et al., 2000], Epstein [Epstein & Gall, 1987], Cremisi [Cremisi et al., 1992], Green [Green et al., 1993] sowie in der sub-viral Datenbank [Rocheleau & Pelchat, 2006]. Diese werden in einem MSA aligniert (Abbildung 2.5).

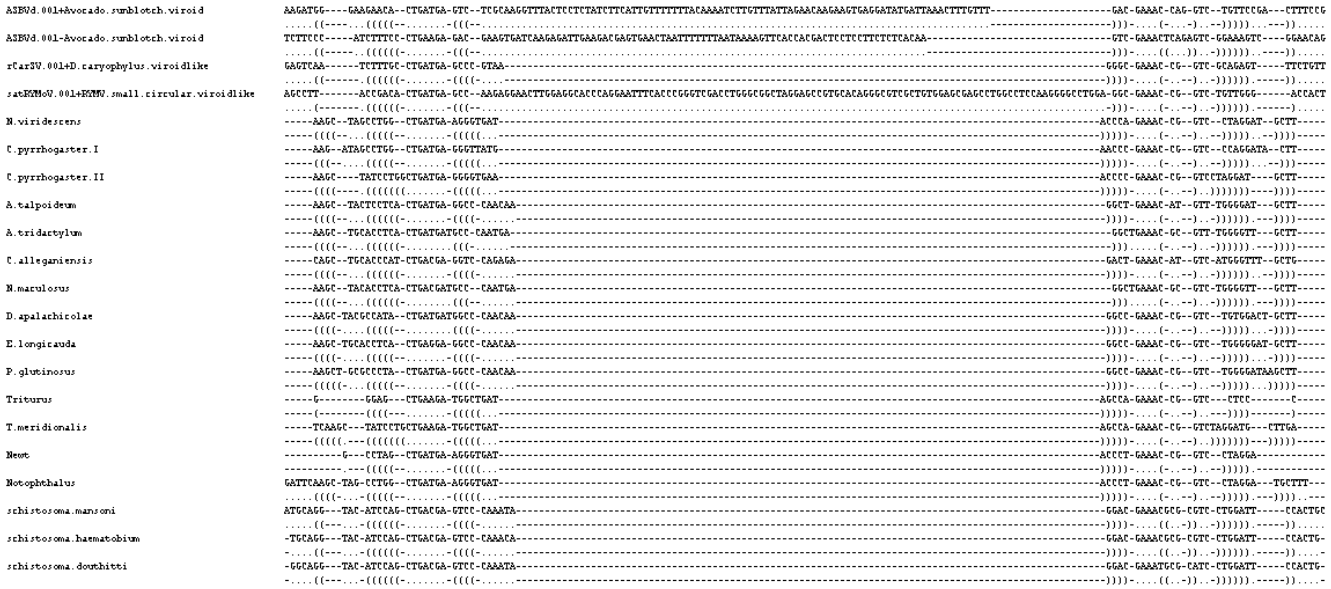


Abbildung 2.5: MSA bekannter Typ I *Hammerhead* Ribozyme
Die bekannten Sequenzen wurden manuell auf Grundlage der Strukturinformation untereinander aligniert und ergeben das dargestellte MSA.

Daraus kann der folgenden Deskriptor abgeleitet werden:

5'-X(D)-X(N)-X(B)-X(H)-H-X(W)-X(M)-X(R)-X(V)-N-X(W)-X(D)-X(V)-N-X(W)-X(H)-B-N-N-N-N-C-T-G-A-N-G-A
-X(T)-D-N-B-X(B)-X(Y)-N-H-D-X[104]-X(W)-X(R)'-X(V)'-V'-N'-H'-X(T)-G-A-A-A-Y-X(K)-X(C)-V-B-X(M)'-R'
-T'-Y-N'-N'-N'-N'-V'-X(D)'-X(D)'-X(N)-X(N)-X(W)-N'-X(B)'-X(Y)'-X(K)'-X(W)'-H-X(H)-X(N)-X(N)-X(B)-3'

oder anders dargestellt in Abbildung 2.6.

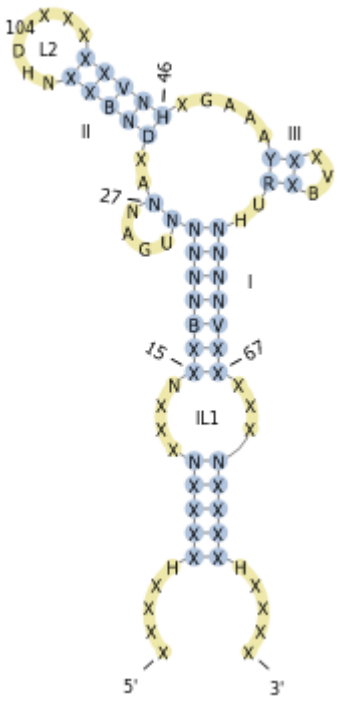


Abbildung 2.6: *Hammerhead* Ribozym Typ I Konsensussequenz
2.6 ist die Konsensussequenz aus 21 bekannten HHRz I aus 2.5. X entspricht einem möglichen zusätzlichen Nukleotid. *Loop* II kann eine Größe von 3 bis 110 nt besitzen.

X entspricht möglichen zusätzlichen Positionen und X(B) z. B. den beobachteten Basen (C, T, G). Bezüglich der Suchlaufzeit mit PatScan sollten Definitionen, wie „DNBX“, wenn möglich allgemeiner formuliert werden, also als „NNNX“, da ein genauer Zeichenvergleich zeitintensiv ist. Der letztendlich verwendete Deskriptor ist in Abbildung 2.7(d) zu sehen.

Motivsuche

Die einzelnen Ergebnisse der Motivsuchen sind auf dem Server (IP-Adresse: 141.50.190.146) in einer MySQL Datenbank und im Dateisystem unter „/media/sdb2009/src/“ abgelegt. Je nach Quelle der Daten gibt es einen „genomes“ Ordner, z. B. „ensembl-genomes“, der verschiedene Unterordner besitzt. Jeder Unterordner entspricht einem Organismus und enthält je nach *Assembly* gepackte Dateien der Chromosomen bzw. *Contigs* oder *Scaffolds*.

Das Datum ist neben dem Namen des Suchmusters und des Organismus Bestandteil des Dateinamens. Alle Dateien einer Suche haben das gleiche Datum. Zum Beispiel die Datei

```
/media/sdb2009/src/ensembl-genomes/xenopus_tropicalis/
xenopus_tropicalis_hammerhead_typI_ugan_gaaa_uh_hitlist_7Jan2011_4682.txt
```

enthält das Ergebnis der Suche vom 07.01.2011 nach *Hammerhead* Ribozymen vom Typ I mit dem konservierten, katalytischen Zentrum UGAN, GAAA und UH aus dem Genom *Xenopus tropicalis*, welches ein Organismus der Ensembl Datenbank ist. Die letzte Zahl ist ein interner Identifizierer der Genomdateien.

Für die verschiedenen Motivsuchen dieser Arbeit wurden auf Sekundärstrukturen basierende Deskriptoren verwendet. Die Abbildungen 2.7, 2.8, 2.9, 2.10, 2.11, und 2.12 zeigen die verschiedenen Deskriptoren und die Tabellen 2.1, 2.2 und 2.3 eine Übersicht der durchgeführten Suchen sowie Parameter- und Filtereinstellungen.

2

Tabelle 2.1: Motivsuchen

Datum	Suche	Faltung	Name
10. Jul. 2009	PatScan	UNAfold	HHRzIII
03. Sep. 2009	PatScan	nein	HHRzIII
03. Sep. 2009	PatScan	nein	HHRzII

Tabelle 2.1 – Fortsetzung

Datum	Suche	Faltung	Name
03. Sep. 2009	PatScan	nein	HHRzI
14. Sep. 2009	RNAbob	UNAfold	HHRzII
16. Sep. 2009	PatScan	UNAfold	HHRzII
16. Sep. 2009	PatScan	UNAfold	HHRzIII
01. Okt. 2009	RNAbob	UNAfold	HHRzI
26. Okt. 2009	PatScan	nein	HairpinRz
28. Okt. 2009	RNAbob	Mfold	HHRzII
10. Nov. 2009	RNAbob	Mfold	HHRzIII
02. Dez. 2009	RNAbob	Mfold	HHRzII
11. Jan. 2010	RNAbob	nein	HairpinRz
14. Jan. 2010	RNAbob	Mfold	HairpinRz
14. Jan. 2010	RNAbob	Mfold	HHRzII
17. Mär. 2010	RNAbob	Mfold	HairpinRz
19. Mär. 2010	RNAbob	nein	HHRzII
22. Mär. 2010	RNAbob	nein	HHRzI
26. Mär. 2010	RNAbob	Mfold	HHRzIII
16. Jul. 2010	RNAbob	Mfold	HHRzIII
05. Aug. 2010	PatScan	Mfold	HHRzIII
16. Aug. 2010	PatScan	Mfold	HHRzIII
27. Aug. 2010	RNAbob	Mfold	HHRzIII
02. Sep. 2010	PatScan	Mfold	HHRzIII
28. Sep. 2010	PatScan	UNAfold	HHRzIII
04. Okt. 2010	PatScan	Mfold	HHRzIII
03. Nov. 2010	RNAbob	Mfold	HHRzIII
03. Nov. 2010	PatScan	Mfold	HHRzII
12. Nov. 2010	PatScan	Mfold	HHRzIII
03. Dez. 2010	PatScan	Mfold	HHRzIII
20. Dez. 2010	PatScan	Mfold	HHRzI
07. Jan. 2011	PatScan	Mfold	HHRzI
24. Feb. 2011	PatScan	nein	HDVRz
29. Apr. 2011	PatScan	Mfold	HHRzIII
18. Jul. 2011	PatScan	Mfold	HHRzIII

[2] Übersicht der nach dem Datum sortierten Suchen mit den jeweiligen Such- und Faltungsprogrammen sowie den Motivnamen. Die einzelnen Suchen wurden nicht zwingend auf den Gesamtdaten ausgeführt.

Tabelle 2.2 zeigt die jeweiligen Faltungseinschränkungen, die bei der Formulierung der *Constraints* zur Berechnung der freien Energie einer erzwungen Struktur verwendet wurden.

[3]

Tabelle 2.2: Faltungsparameter

Datum	Einschränkung
10. Jul. 2009	keine
03. Sep. 2009	keine
03. Sep. 2009	keine
03. Sep. 2009	keine
14. Sep. 2009	keine
16. Sep. 2009	keine
16. Sep. 2009	keine
01. Okt. 2009	keine
26. Okt. 2009	keine

Tabelle 2.2 – Fortsetzung

Datum	Einschränkung
28. Okt. 2009	keine
10. Nov. 2009	h3=N s3=UGAN h3'=N s4=A s6=GAA
02. Dez. 2009	s2=GAA h3=N h3'=N h5=N s6=UGAN h5'=N s7=A
11. Jan. 2010	keine
14. Jan. 2010	keine
14. Jan. 2010	s2=GAA h3=N h3'=N h5=N s6=UGAN h5'=N s7=A
17. Mär. 2010	keine
19. Mär. 2010	keine
22. Mär. 2010	keine
26. Mär. 2010	h3=N s3=UGAN h3'=N s4=A s6=GAA
16. Jul. 2010	h3=N s3=UGAN h3'=N s4=A s6=GAA
05. Aug. 2010	p3=N s3=UGAN p3'=N s4=A s6=GAA
16. Aug. 2010	s3=CUGANGA s5=GAA
27. Aug. 2010	s3=CUGANGA s5=GAA
02. Sep. 2010	s3=CUGANGA s5=GAA
28. Sep. 2010	p3=N s3=UGAN p3'=N s4=A s6=GAA
04. Okt. 2010	s3=CUGANGA s5=GAA
03. Nov. 2010	s3=CUGANGA s5=GAA
03. Nov. 2010	p=1 s4=GAA s9=CUGANGA
12. Nov. 2010	s3=CUGANGA s5=GAA
03. Dez. 2010	s3=CUGANGA s5=GAA
20. Dez. 2010	p1=1 s3=CUGANGA s5=GAA p4=2
07. Jan. 2011	p1=1 p3=1 s3=UGAN s4=A s6=GAA p5=2
24. Feb. 2011	keine
29. Apr. 2011	p3=1 s3=UGAN s4=A s6=GAA
18. Jul. 2011	p3=1 s3=UGAN s4=A s6=GAA

[3] Übersicht der nach dem Datum sortierten Suchen mit den jeweiligen Faltungseinschränkungen, aufgrund derer die freie Energie einer erzwungenen Struktur berechnet wurde.

Die verwendeten Filtereinstellungen, welche aufgrund entsprechender Laborergebnisse angepasst wurden, werden in Tabelle 2.3 zusammengefasst.

[4]

Tabelle 2.3: Filtereinstellungen

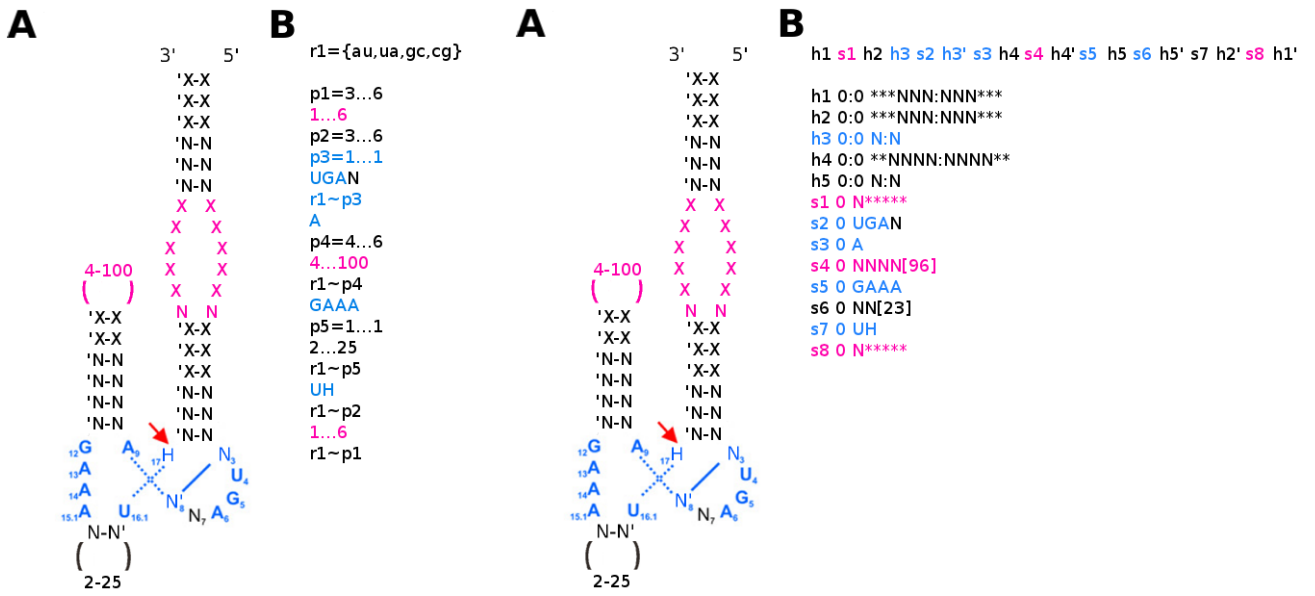
Datum	ΔG_{free}	$\Delta\Delta G$	Überlappend	Wobble-bp	Unique
10. Jul. 2009	-	$0.6\frac{kcal}{mol}$	±5	-	+
03. Sep. 2009	-	-	-	-	-
03. Sep. 2009	-	-	-	-	-
03. Sep. 2009	-	-	-	-	-
14. Sep. 2009	-	$0.6\frac{kcal}{mol}$	±5	-	+
16. Sep. 2009	-	$0.6\frac{kcal}{mol}$	±5	-	+
16. Sep. 2009	-	$0.6\frac{kcal}{mol}$	±5	-	+
01. Okt. 2009	-	$0.6\frac{kcal}{mol}$	±5	-	+
26. Okt. 2009	-	-	-	-	-
28. Okt. 2009	-	$0.6\frac{kcal}{mol}$	±5	-	+
10. Nov. 2009	-	$0.6\frac{kcal}{mol}$	±5	-	+
02. Dez. 2009	-	$0.6\frac{kcal}{mol}$	±5	-	+
11. Jan. 2010	-	-	-	-	-
14. Jan. 2010	-	$0.6\frac{kcal}{mol}$	±5	-	+
14. Jan. 2010	-	$0.6\frac{kcal}{mol}$	±5	-	+
17. Mär. 2010	-	$0.6\frac{kcal}{mol}$	±5	-	+
19. Mär. 2010	-	-	-	-	-

Tabelle 2.3 – Fortsetzung

Datum	ΔG_{free}	$\Delta\Delta G$	Überlappend	Wobble-bp	Unique
22. Mär. 2010	-	-	-	-	-
26. Mär. 2010	$-10\frac{kcal}{mol}$	$0.5\frac{kcal}{mol}$	±5	-	+
16. Jul. 2010	$-14\frac{kcal}{mol}$	$0.6\frac{kcal}{mol}$	±5	-	+
05. Aug. 2010	$-14\frac{kcal}{mol}$	$0.6\frac{kcal}{mol}$	±5	-	+
16. Aug. 2010	$-10\frac{kcal}{mol}$	$1\frac{kcal}{mol}$	±5	-	+
27. Aug. 2010	-	$0\frac{kcal}{mol}$	±5	+	+
02. Sep. 2010	$-10\frac{kcal}{mol}$	$0\frac{kcal}{mol}$	±5	-	+
28. Sep. 2010	$-10\frac{kcal}{mol}$	$1\frac{kcal}{mol}$	±5	-	+
04. Okt. 2010	$-10\frac{kcal}{mol}$	$1\frac{kcal}{mol}$	±5	-	+
03. Nov. 2010	$-10\frac{kcal}{mol}$	$0.5\frac{kcal}{mol}$	±5	-	+
03. Nov. 2010	$-10\frac{kcal}{mol}$	$0.5\frac{kcal}{mol}$	±5	-	+
12. Nov. 2010	$-10\frac{kcal}{mol}$	$0.5\frac{kcal}{mol}$	±5	-	+
03. Dez. 2010	$-10\frac{kcal}{mol}$	$0.5\frac{kcal}{mol}$	±5	-	+
20. Dez. 2010	$-10\frac{kcal}{mol}$	$0.5\frac{kcal}{mol}$	±6	-	+
07. Jan. 2011	$-7\frac{kcal}{mol}$	$2\frac{kcal}{mol}$	±6	-	+
24. Feb. 2011	-	-	-	-	-
29. Apr. 2011	$-10\frac{kcal}{mol}$	$0.5\frac{kcal}{mol}$	±5	+	+
18. Jul. 2011	$-14\frac{kcal}{mol}$	$0\frac{kcal}{mol}$	±5	+	+

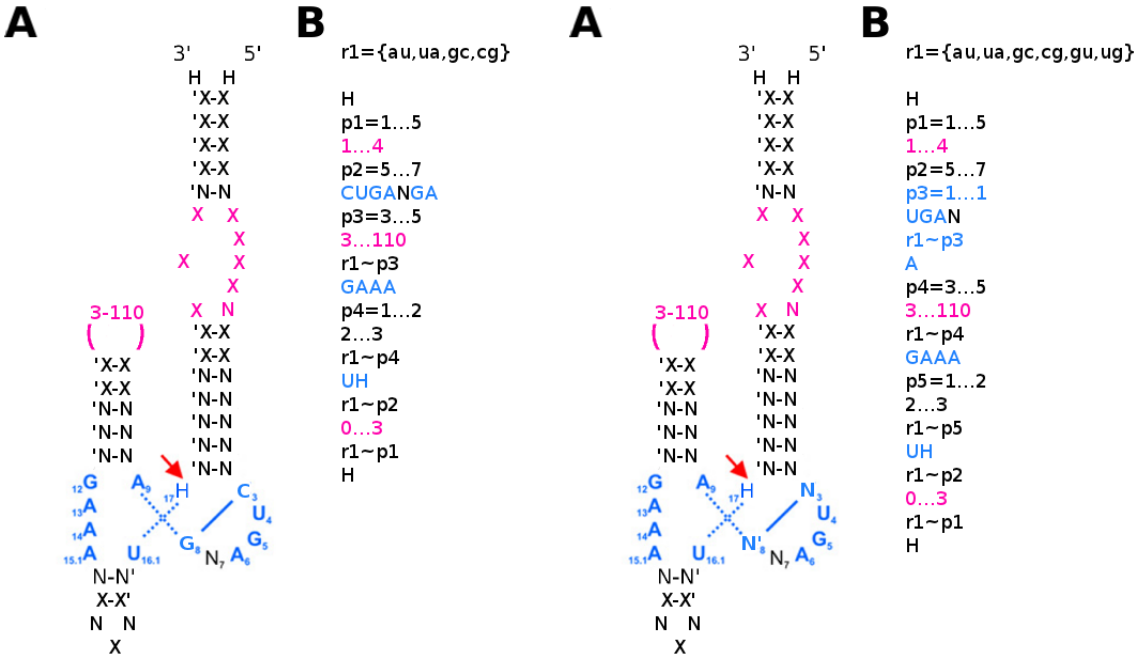
[4] Übersicht der nach dem Datum sortierten Suchen mit den jeweiligen Filtereinstellungen. Der $\Delta\Delta G$ Wert entspricht vor dem 16.08.2010 dem Quotienten aus ΔG_{motif} und ΔG_{free} und danach der Differenz.

Alle Suchen wurden überlappend durchgeführt, da verschachtelte Motive beobachtet wurden, die durch den modularen Aufbau innerhalb einer Region vorkommen können.



(a) 03.09.2009

(b) 01.10.2009 und 22.03.2010

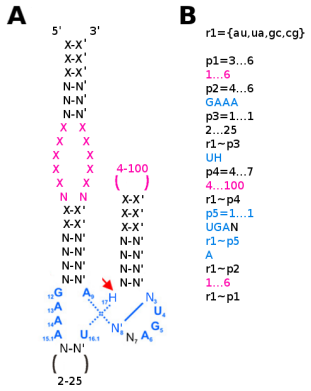


(c) 20.12.2010

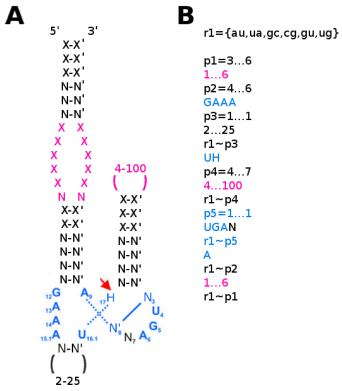
(d) 07.01.2011

Abbildung 2.7: *Hammerhead* Ribozym Typ I Deskriptoren

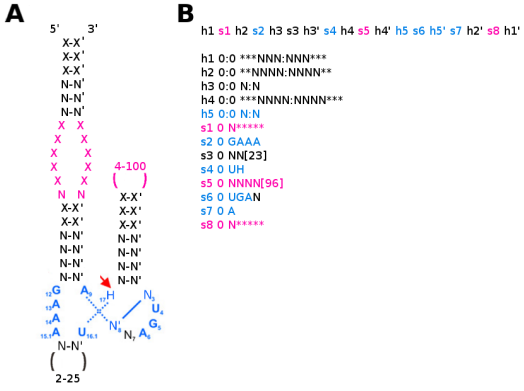
A) zeigt die Sekundärstruktur der HHRz I Deskriptoren. Der rote Pfeil markiert die Spaltstelle. In blau sind alle konservierten Nukleotide des katalytischen Zentrums dargestellt und gemäß [Hertel et al., 1992] durchnummeriert. Alternative Positionen sind durch ein X gekennzeichnet. X' bzw. N' symbolisieren eine komplementäre Base. Die interagierenden *Loops* sind in mangenta gefärbt. B) zeigt die Übersetzung in die jeweilige Syntax des Suchprogrammes (siehe Abschnitt 2.4.12). (a), (b) besitzen den gleichen Deskriptor, mit verschiedenen Syntaxen (PatScan (a), RNAbob (b)) und (c), (d) modifizierte Deskriptoren mit PatScan Definitionen, in denen ausschließlich Watson-Crick Basenpaare und C3G8 (c) sowie Wobble Basenpaare und N3N8 (d) zugelassen wurden.



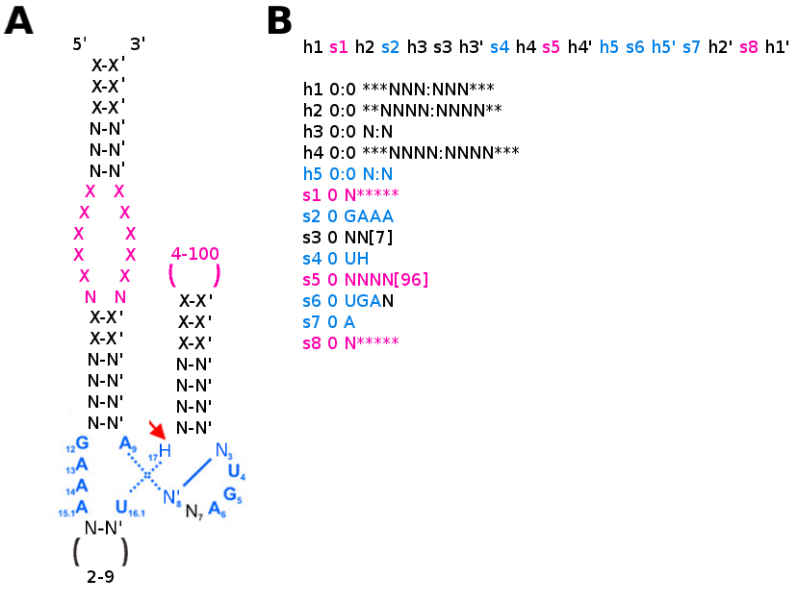
(a) 03.09.2009



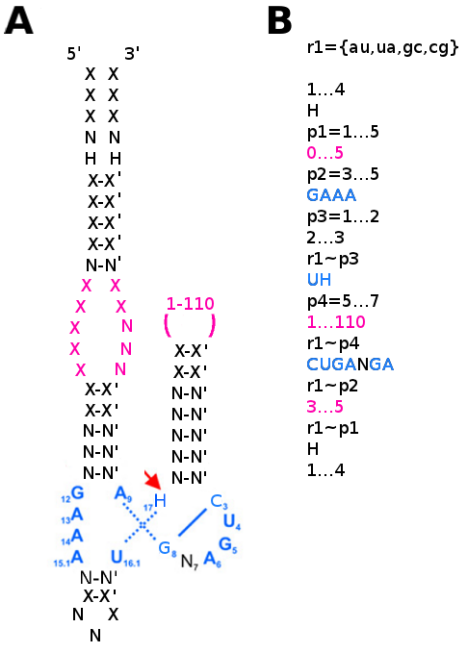
(b) 16.09.2009



(c) 14.09, 28.10, 02.12.2009 und 19.03.2010



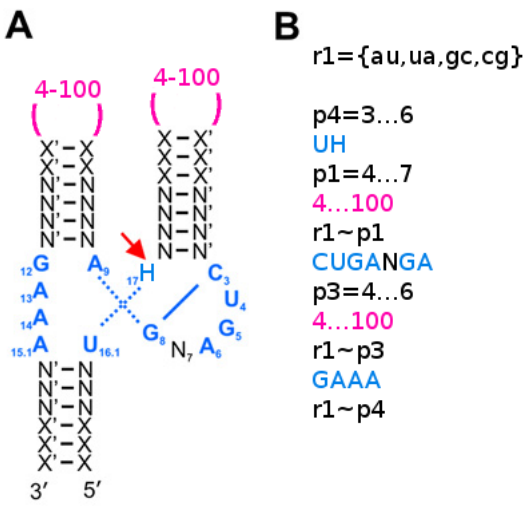
(d) 14.01.2010



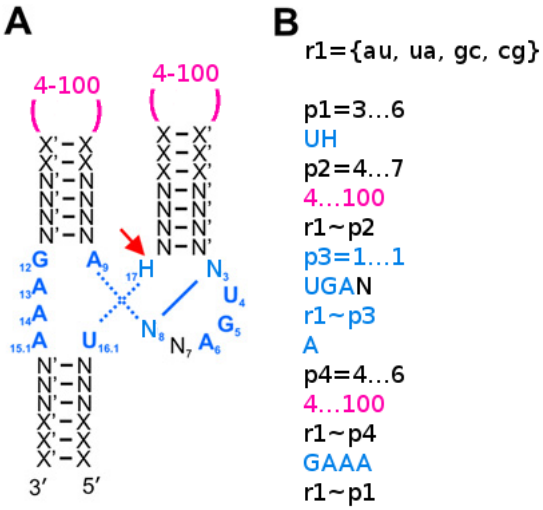
(e) 3.11.2010

Abbildung 2.8: *Hammerhead* Ribozym Typ II Deskriptoren

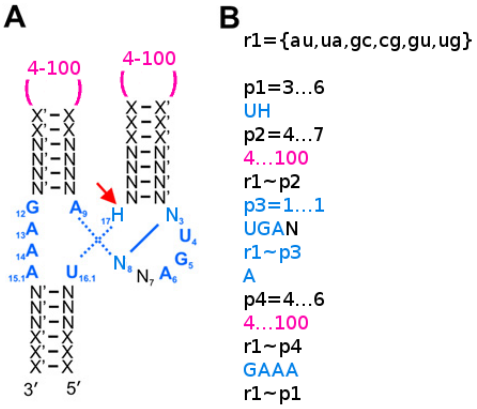
In A) sind die HHRz II Sekundärstrukturen zu sehen. Die Spaltstelle ist durch den roten Pfeil markiert und in blau sind alle konservierten Nukleotide des katalytischen Zentrums dargestellt. X entspricht einem möglichen zusätzlichen Nukleotid. X' bzw. N' symbolisieren eine komplementäre Base. Die interagierenden *Loops* sind in mangenta gefärbt. B) zeigt die PatScan bzw. RNAbob Definition. Der Deskriptor in (a), (b) und (c) ist der Gleiche, jedoch mit unterschiedlicher *Pattern*-Beschreibung (PatScan (a), (b), RNAbob (c)). In (d) wurde *Loop* 3 verkleinert. (e) besitzt die meisten Änderungen.



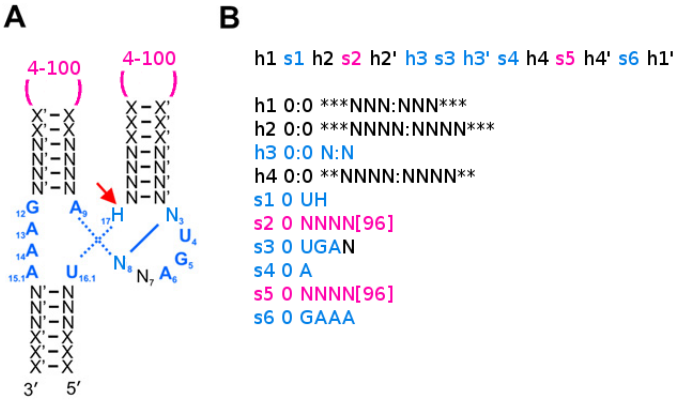
(a) 10.07.2009



(b) 03.09.2009, 05.08 und 28.09.2010



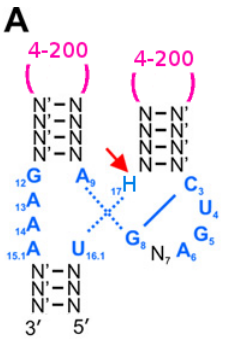
(c) 16.09.2009, 29.04. und 18.07.2011



(d) 10.11.2009, 26.03 und 16.07.2010

Abbildung 2.9: *Hammerhead* Ribozym Typ III Deskriptoren

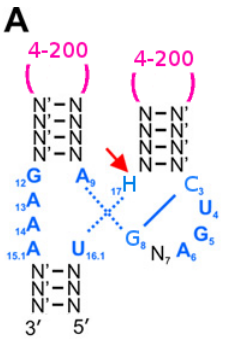
A) zeigt die Sekundärstruktur der HHRz III Motive. Der rote Pfeil markiert die Spaltstelle und in blau sind alle konservierten Nukleotide des katalytischen Zentrums dargestellt. X kennzeichnet mögliche zusätzliche Positionen. X' bzw. N' symbolisieren eine komplementäre Base. Interagierende *Loops* sind in mangenta gefärbt. B) zeigt die Übersetzung in die Syntax des Suchprogrammes. (b), (c) und (d) besitzen den gleichen Deskriptor. Der Unterschied zum Deskriptor in (a) ist das C3G8 Basenpaar. In (a), (b) und (c) wurde mit PatScan gesucht, in (d) mit RNAbob, wobei (a), (b) ausschließlich Watson-Crick Basenpaarungen erlauben und (c), (d) Wobble Basenpaarungen.



B

```
r1={au,ua,gc,cg}

p1=3...3
UH
p2=4...4
4...200
r1~p2
CUGANGA
p3=4...4
4...200
r1~p3
GAAA
r1~p1
```

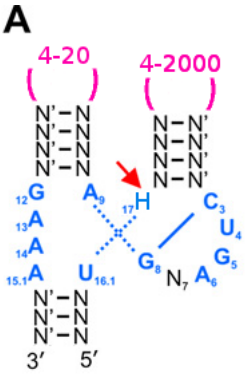
(a) 16.08. und 02.09.2010



B

```
h1 s1 h2 s2 h2' s3 h3 s4 h3' s5 h1'

h1 0:0 NNN:NNN
h2 0:0 NNNN:NNNN
h3 0:0 NNNN:NNNN
s1 0 UH
s2 0 NNNN[196]
s3 0 CUGANGA
s4 0 NNNN[196]
s5 0 GAAA
```

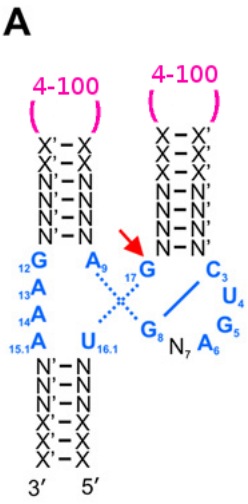
(b) 27.08.2010



B

```
r1={au,ua,gc,cg}

p1=3...3
UH
p2=4...4
4...2000
r1~p2
CUGANGA
p3=4...4
4...20
r1~p3
GAAA
r1~p1
```

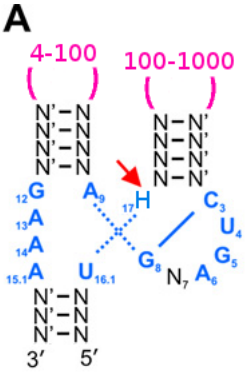
(c) 04.10.2010



B

```
h1 s1 h2 s2 h2' s3 h3 s4 h3' s5 h1'

h1 0:0 ***NNN:NNN***
h2 0:0 ***NNNN:NNNN***
h3 0:0 **NNNN:NNNN**
s1 0 UG
s2 0 NNNN[96]
s3 0 CUGANGA
s4 0 NNNN[96]
s5 0 GAAA
```

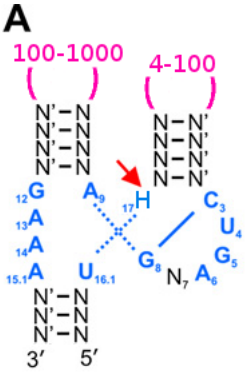
(d) 03.11.2010



B

```
r1={au,ua,gc,cg}

p1=3...3
UH
p2=4...4
100...1000
r1~p2
CUGANGA
p3=4...4
4...100
r1~p3
GAAA
r1~p1
```

(e) 12.11.2010



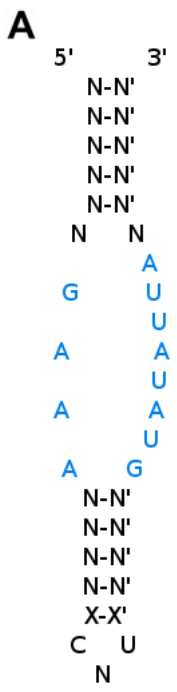
B

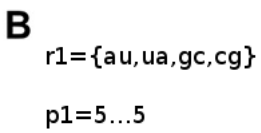
```
r1={au,ua,gc,cg}

p1=3...3
UH
p2=4...4
4...100
r1~p2
CUGANGA
p3=4...4
100...1000
r1~p3
GAAA
r1~p1
```

(f) 03.12.2010

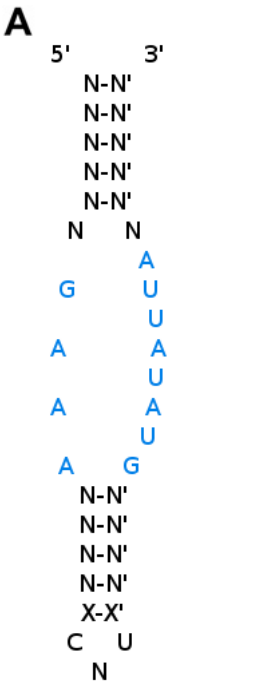
Abbildung 2.10: *Hammerhead* Ribozym Typ III Deskriptoren (Fortsetzung)

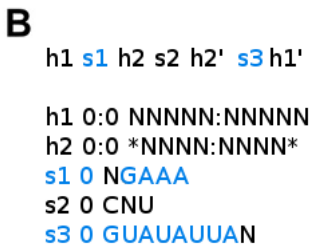
Weitere HHRz III Deskriptoren mit den jeweiligen in A) dargestellten Sekundärstrukturen. Der rote Pfeil markiert die Spaltstelle und in blau sind alle konservierten Nukleotide des katalytischen Zentrums zu sehen. X entspricht einem möglichen zusätzlichen Nukleotid und X' bzw. N' symbolisieren komplementäre Basen. Interagierende *Loops* sind in mangenta dargestellt. B) zeigt die Deskriptordefinition des Suchprogrammes. (a) und (b) besitzen den gleichen Deskriptor mit unterschiedlichen Definitionen für PatScan (a) und RNAbob (b). In (c), (e) und (f) wurden die *Loop*-Größen verändert und in (d) mit RNAbob nach U16.1G17 gesucht.



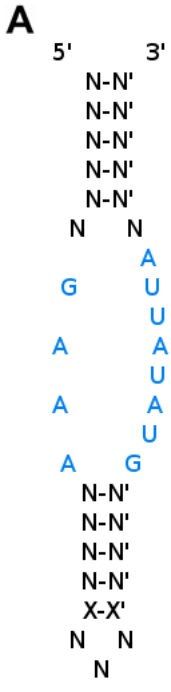


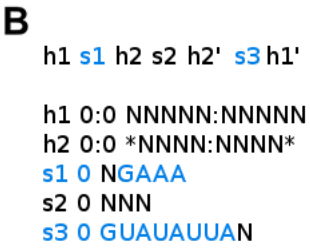
(a) 26.10.2009



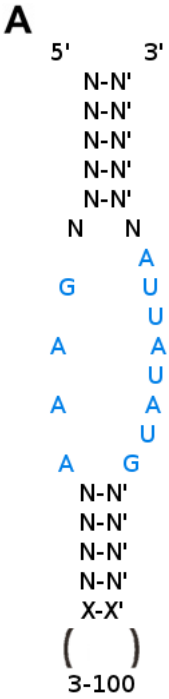


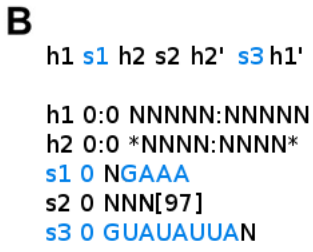
(b) 11.01.2010





(c) 14.01.2010

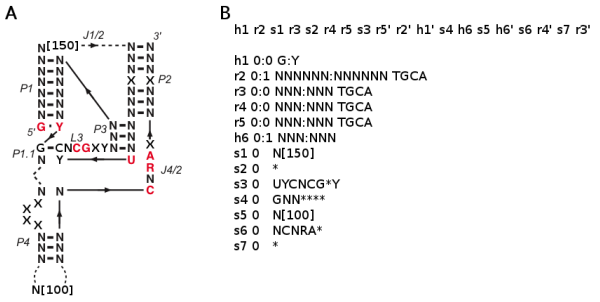




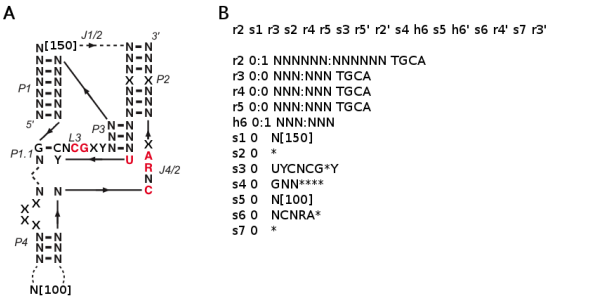
(d) 17.03.2010

Abbildung 2.11: *Hairpin* Ribozym Deskriptoren

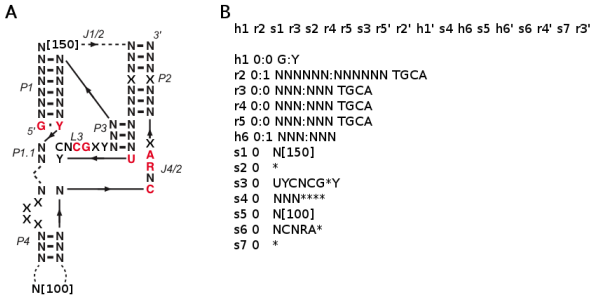
A) zeigt einen Teil der Sekundärstruktur des *Hairpin* Ribozyms und B) die Übersetzung in die jeweilige Syntax des Suchprogrammes. Blau gefärbte Nukleotide sind in den *Hairpin* Ribozymen konserviert. (a) und (b) enthalten den gleichen Deskriptor mit verschiedener Motivbeschreibung für PatScan (a) und RNAbob (b). In (c) wurde der *Loop* verallgemeinert und in (d) erweitert.



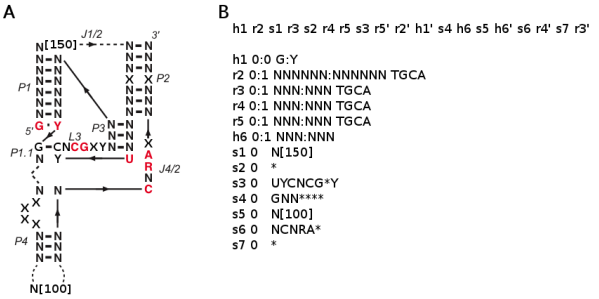
(a) 24.02.2011



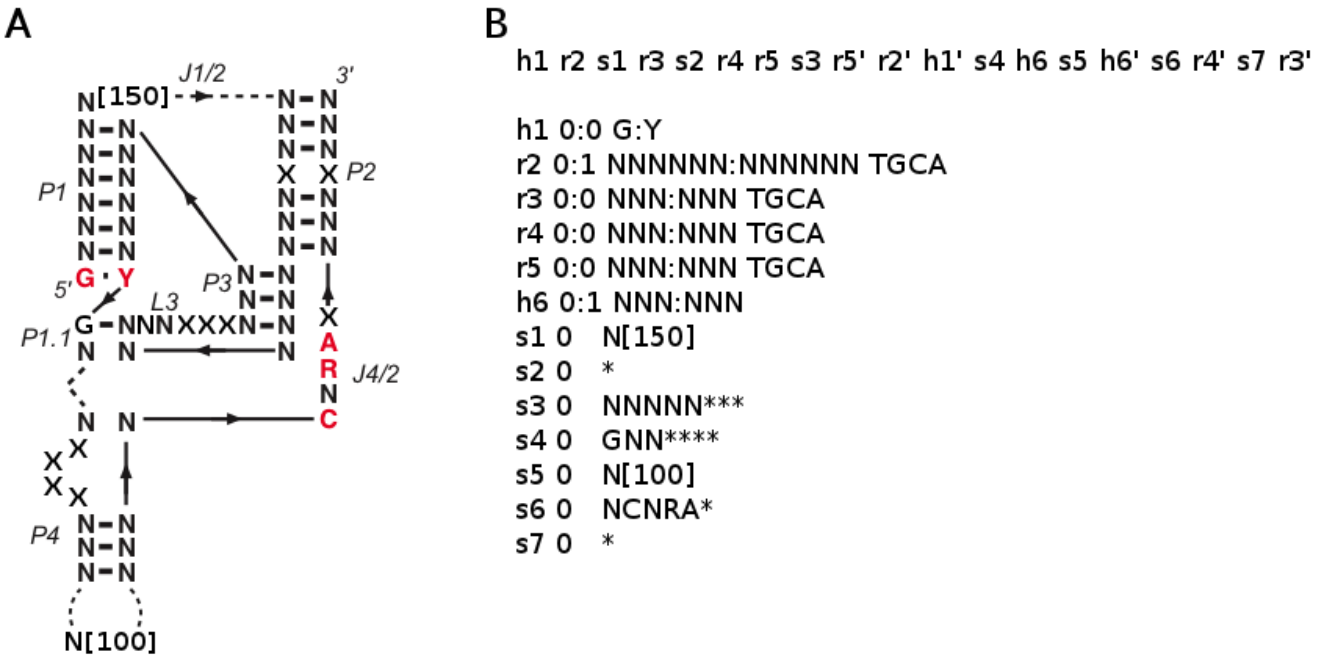
(b) 24.02.2011



(c) 24.02.2011



(d) 24.02.2011



(e) 24.02.2011

Abbildung 2.12: HDV Ribozym Deskriptoren

A) zeigt die aus [Webb et al., 2009] modifizierte Sekundärstruktur eines HDV Ribozyms. P1 bis P4 entsprechen Helices und J1/2, J4/2 Einzelstrangregionen. Mögliche zusätzliche Nukleotide sind durch ein X gekennzeichnet. Y und R entsprechen einem Pyrimidin bzw. Purin gemäß Tabelle 1.1. Rote Nukleotide sind konservierte Positionen. B) zeigt die Übersetzung in die RNAbob Syntax (siehe Abschnitt 2.4.12). Im Vergleich zu (a) aus [Webb et al., 2009] wurde in (b) das GY Basenpaar entfernt, in (c) Helix P1.1 und in (e) L3 verallgemeinert. (d) besitzt den gleichen Deskriptor wie (a), erlaubt jedoch in der Motivbeschreibung in den Helices r3, r4 und r5 *Mismatches*.

2.4.18 Deep Sequencing

Im Rahmen des zweiten Projektes dieser Arbeit wurden die folgenden Techniken und Programme verwendet.

Illumina Deep Sequencing

Die Illumina® Sequenzierung (www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf) gehört zur zweiten Generation der Sequenzierungstechnologien und ermöglicht eine schnelle und genaue Sequenzierung. Zunächst werden die gegebenen Sequenzen an eine Adaptersequenz ligiert und auf einer Oberfläche immobilisiert. Diese werden anschließend über eine Brückenamplifizierung vervielfältigt und bilden nach mehreren Zyklen dichte Cluster aus Millionen Kopien der einzelnen Sequenzen. Die Sequenzierung mittels Synthese (SBS)-Technologie verwendet für jedes Nukleotid eine Fluoreszenzmarkierung, die von einem Laser erkannt und deren Signalstärke gemessen wird. Die Signalstärke ist für eine spätere Bewertung der Sequenzierungsqualität nützlich (Phred Quality Score). Die Fluoreszenzmarkierung dient als Terminator für die DNA Polymerase und wird nach der Erfassung am Ende eines Zyklus entfernt, um das nächste Nukleotid einbauen zu können. Somit ergeben sich nach und nach die Nukleotide der einzelnen Sequenzpositionen des Clusters. Die Sequenz eines Clusters entspricht einem *Read*.

Illumina Deep Sequencing diente der Sequenzierung kleiner RNA aus einem *Dictyostelium discoideum* AX2 Wildtyp und einem *rrpC* Gendeletionsstamm, jeweils in zweifacher Ausführung. Beim *rrpC* Gendeletionsstamm wurde das für RrpC kodierende Gen *rrpC* durch homologe Rekombination entfernt [Wiegand et al., 2011].

FastQC

Einer der ersten Schritte bei der Auswertung der Sequenzierungsergebnisse ist die Betrachtung der Sequenzierungsqualität. Dazu dient das Programm FastQC Version 0.7.2 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), welches einen Bericht über die Ergebnisse erzeugt. Es besteht aus verschiedenen Modulen, die helfen, potenzielle Probleme in den Daten zu identifizieren. Zu den Modulen zählen u. a.:

- die Sequenzierungsqualität pro Nukleotid und der Phred Quality Score aller Sequenzen,
- die Nukleotidhäufigkeit pro Sequenzposition,
- der GC-Gehalt pro Sequenzposition und über alle Sequenzen,
- die Sequenzlängenverteilung,
- die Häufigkeit von Duplikationen unter den Sequenzen
- sowie überrepräsentierte Sequenzen und Teilsequenzen.

Auftretende Probleme können entweder während der Sequenzierung entstanden sein oder ihren Ursprung im Ausgangsmaterial haben.

Cutadapt

Die Illumina Sequenzierung verwendet eine Adaptersequenz, die Teil der sequenzierten *Read*-Sequenz sein kann. Ein Programm, das diese entfernen kann, heißt Cutadapt Version 1.0 (http://code.google.com/p/cutadapt/wiki/documentation) [Martin, 2011]. Der verwendete Adapter lautet:

`TruSeq Illumina small RNA 3' adapter RA3 mit der Sequenz (TGGAATTCTCGGGTGCCAAGG).`

Es gibt verschiedene Programmoptionen, die das Erkennen der Adaptersequenz in den *Reads* beeinflussen. Die Standardeinstellung der Fehlerrate (erlaubte *Mismatches* pro Länge) ist 1:10. Für die gegebene Adaptersequenz werden demnach zwei *Mismatches* zugelassen. Ein *Read* muss mindestens 3 nt mit der Adaptersequenz überlappen und wird entsprechend gekürzt.

Bowtie

Der nächste Schritt im Rahmen der Auswertung der überarbeiteten *Reads* ist die Beantwortung der Frage nach der Lokalisation der *Reads* und die damit verbundene Abbildung (*Mapping*) gegen eine Referenzsequenz. Dazu eignet sich Bowtie Version 2.0.0 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) [Langmead et al., 2009]. Bowtie ist ein schnelles und speichereffizientes Programm, das mit Hilfe globaler, lokaler und *paired-end Alignment* Algorithmen kurze *Read*-Sequenzen gegen lange Genomsequenzen alignieren kann. Dazu muss die Genomsequenz zunächst indexiert werden, ähnlich wie bei BLAST. Die Schwierigkeit des *Mappings* liegt darin, abzuschätzen, wo ein *Read* seinen Ursprung hat. Dies ist jedoch aufgrund repetitiver Regionen nicht in jedem Fall bestimmbar. Deshalb bietet Bowtie eine *Mapping Quality* $Q = -10 * log_{10}(Wahrscheinlichkeit(Falsche\ Position))$, um die Glaubwürdigkeit des Ursprungs eines *Reads* abschätzen zu können. Sie kann als eine Art Einzigartigkeit des *Reads* betrachtet werden. Ein *Alignment* ist einzigartig, wenn der Score größer ist als bei allen anderen *Alignments*. Je größer der Abstand der Scores zwischen dem besten und zweit besten *Alignment*, desto größer ist die Einzigartigkeit. Eine *Mapping Quality* von 10 oder kleiner bedeutet, dass es eine 1 zu 10 Chance gibt, dass das *Read* einen anderen Ursprung besitzt.

Als Referenzsequenz diente zum einen das *Dictyostelium discoideum* Genom heruntergeladen von http://dictybase.org/ und zum anderen eine Bibliothek aus eigenen *Repeat*-Sequenzen von Johan Reimegård. Darin enthalten sind einfache *Repeats* aus Repbase [Jurka et al., 2005], wie $(CAT)_n$ n mal hintereinander und Konsensussequenzen komplexer *Repeats* wie die Retrotransposons DIRS-1 und Skipper. Ein Teil der Bibliothek befindet sich im Anhang. Die *Repeats* aus Repbase können nach erhaltener Lizenz vom Server http://www.girinst.org/server/RepBase/index.php heruntergeladen werden.

Es gibt viele Optionen, die das *Scoring, Mapping* und die Ausgabe der *Reads* beeinflussen. Die Standardeinstellung ist ein globales (*end-to-end*) *Alignment* mit einem Mindestscore von $(-0.6 + -0.6 * Read\ Länge)$, um als „gültiges *Alignment*“ zu zählen. Der Score berechnet sich aus der Summe der Strafwerte jedes Unterschiedes (Mismatch = -6, Gap open = -5, Gap extension = -3). Vor dem *Mapping* wird jede *Read*-Sequenz und die dazugehörige reverskomplementäre Sequenz in überlappende Teilsequenzen (*Seeds*) aufgeteilt. Die standard *Seed*-Länge beträgt 20 nt ohne *Mismatches* mit einem $f(x) = 1 + 1.25 * \sqrt{(Read\ Länge)}$ *Shift* zwischen den Teilsequenzen. Das Beste aus vier gültigen, nichtüberlappenden *Alignments* wird im SAM Format ausgegeben. Es wurden mehrere unterschiedliche Einstellungen getestet (siehe Tabelle 3.10).

- -M x das Beste aus x+1 gültigen *Alignments* wird pro *Read* ausgegeben
- -D x Anzahl dynamischer Programmierungsprobleme (z. B. *Seed*-Erweiterung), die misslingen können, bevor die Suche nach dem besten *Alignment* beendet wird
- -R x maximale Anzahl an „re-seed“ Versuchen um repetitive *Seeds* zu alignieren
- -k x pro *Read* werden x gültige *Alignments* absteigend sortiert nach dem Score ausgegeben
- -N x Anzahl *Mismatches* pro *Seed*
- -L x setzt die *Seed*-Länge
- -i setzt das Intervall zwischen den extrahierten *Seeds*

Weitere Optionen und Erklärungen befinden sich im Bowtie2 Handbuch (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml).

SAMtools

SAMtools Version 0.1.18 (http://samtools.sourceforge.net/) [Li et al., 2009] ist ein Programmpaket, das die Manipulierung von SAM Dateien ermöglicht. Dazu zählen u. a. das Konvertieren, Sortieren und Indexieren von SAM und *Binary Alignment Map* (BAM) Dateien.

HTseq

HTseq Version 0.5.1 (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html) ist ein Programm zum Ermitteln von *Read*-Häufigkeiten. Es zählt die *Reads*, die eine bestimmte annotierte Region überlappen.

Integrative Genomics Viewer

Ein Programm zur Visualisierung der *Mapping*-Ergebnisse ist z. B. der *Integrative Genomics Viewer* (IGV) Version 2.0.30 [Robinson et al., 2011]. Der Vorteil des Programmes besteht darin, dass weitere Referenzsequenzen geladen werden können. Auf diese Weise konnten alle *Reads*, die gegen das *Dictyostelium discoideum* Genom und die *Repeat*-Bibliothek aligniert wurden, dargestellt werden (siehe Abschnitt 3.3). Eine ausführliche Programmbeschreibung befindet sich auf der Webseite (http://www.broadinstitute.org/software/igv/home). Andere getestete Programme, wie z. B. Savant (http://genomesavant.com/savant/) hatten aufgrund der großen *Read*-Anzahl Speicherprobleme und konnten die *Alignments* nicht darstellen.

Cufflinks

Eine weitere Frage die sich stellt ist, von welchen Genen zwischen Wildtyp und Gendeletionsstamm signifikant unterschiedliche Mengen kleiner RNAs exprimiert werden. Dazu wurde das Programmpaket Cufflinks Version 1.1.0 (http://cufflinks.cbcb.umd.edu/) [Trapnell et al., 2010] verwendet. Es ermöglicht zwei Vorgehensweisen. Zum einen ist Cufflinks ein Werkzeug zur Transkriptom Annotation und zur Abschätzung der Isoformquantifizierung von RNA-Seq *Reads*. Es verwendet dabei das *Mapping*-Ergebnis (sortierte BAM Dateien) und erstellt einen Datensatz aus wahrscheinlichen Transkripten unter Berücksichtigung eventueller Bias in den Ausgangsdaten. Zum anderen ermöglicht Cuffdiff die Identifizierung differentiell expremierter und regulierter Gene. Es wurden verschiedene Programmoptionen getestet. Mit der Option -g kann Cufflinks geführt mit Hilfe einer gegebenen Annotation aufgerufen werden. Aus dieser werden Transkripte zusätzlich als künstliche *Reads* dem *Assembly* hinzugefügt. In der Ausgabe sind dann alle bekannten und neuen Transkripte und *Splice*-Varianten enthalten. Eine andere Standardeinstellung, die angepasst werden musste, ist die Option –max-bundle-frags, welche die maximale Fragmentanzahl einer Region festlegt, bevor diese verworfen wird. Der Standardwert liegt bei 1 Million und wurde auf 1 Milliarde erhöht, da sämtliche DIRS-1 *Reads* verworfen wurden. Anschließend wurden die Transkripte mit Hilfe von Cuffcompare zum Vergleich mit der Referenzannotation herangezogen und die Vereinigung aller Replikate in einer Datei zusammengefügt. Diese Datei diente danach Cuffdiff als Annotation.

Weitere Hilfe befindet sich im Handbuch (http://cufflinks.cbcb.umd.edu/manual.html) und ein Beispiel-Skript im Anhang.

DESeq

Eine andere Programmbibliothek zur Bestimmung signifikanter Unterschiede zwischen den Wildtyp- und Gendeletionsstamm-Daten ist DESeq (http://www-huber.embl.de/users/anders/DESeq/) [Anders & Huber, 2010]. Es nutzt dazu die mit HTseq ermittelten *Read*-Häufigkeiten und berechnet daraus den Unterschied.

SLURM

Die Jobverteilung auf dem Cluster im UPPMAX erfolgt über das *Simple Linux Utility for Resource Management* (SLURM). Die sbatch Shell Skripte besitzen zusätzlich im *Header* Jobparameter, wie:

- das Projekt, dem die benötigte Laufzeit abgezogen wird,
- die Art der Partition, ob Kern oder Knoten,
- die Anzahl der Prozessoren
- und die für den Job zur Verfügung gestellte Zeit.

3 Ergebnisse

Das Ziel der Arbeit ist die Identifizierung bekannter und neuer RNA-Motive, mit dem Hauptaugenmerk auf die Ribozyme sowie deren Verbreitung und Analyse. Dazu dient das erstellte Programm RNAhit, in dem verschiedene Programme miteinander zu einer *Pipeline* (Abbildung 3.10) kombiniert werden sowie zahlreiche weitere Skripte. In diesem Kapitel werden die Programme und die dazugehörenden Ergebnisse der verschiedenen Motivsuchen vorgestellt.

Des Weiteren werden die Ergebnisse des zweiten Projektes, der Auswertung von RNA *Deep Sequencing* Daten aus *Dictyostelium discoideum*, präsentiert.

3.1 Skripte

Die Programme, die nachfolgend gezeigt werden, wurden selbst geschrieben und sind daher Teil der Ergebnisse.

3.1.1 Perl

Die Perl-Skripte entsprechen einem Auszug der vorhandenen Skripte, zu finden unter:

```
/media/sda2009/phd/perlscripts/
```

oder im Anhang der Arbeit. Ein Teil dieser Programme wurde in RNAhit übernommen.

Perl-Skripte können sowohl als Datei als auch direkt in der Konsole eingegeben werden. Das folgende Skript dient dem ermitteln der Gesamtgröße aller Sequenzdaten:

```
perl -e 'use File::Basename; open(TA,"<".$ARGV[0]); my $c=0; my $a=0;
while(<TA>){$a++; my $r=$_; chomp($r); chdir(dirname($r));
print "$a.\tcd ".dirname($r)."\n"; print "\tgunzip $r\n"; qx(gunzip $r);
my $t=substr($r,0,-3); my $v=(-s $t); $c+=$v; print "\tgzip $t\n\t\t$c\n";
qx(gzip $t);} close(TA);' ls_of_gz.txt > size_11022011.txt
```

RNAhit

Das „Haupt“-Programm, das im Rahmen dieser Arbeit implementiert wurde, heißt RNAhit. Es befindet sich aktuell in Version 1.23.1. Die Dokumentationen der einzelnen Funktionen befinden sich im Quellcode und können z. B. mit Hilfe des Befehls

```
perldoc RNAhit_1_23_1.pl
```

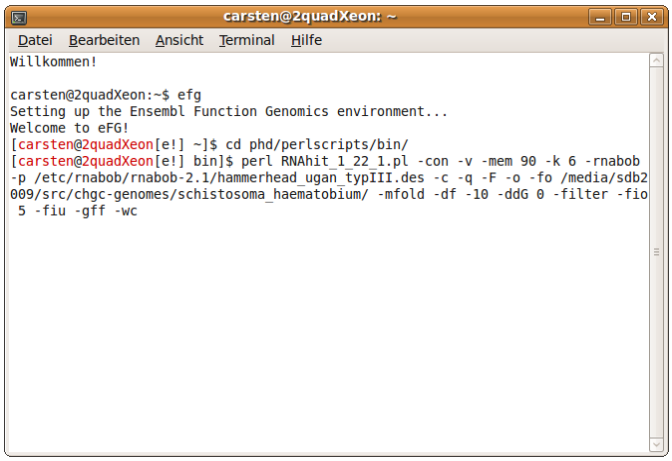
angezeigt werden. Im Quellcode ebenfalls enthalten ist die Historie der einzelnen Programmversionen.

RNAhit kann auf zwei Weisen gestartet werden. Entweder auf Konsolenebene, bei der die einzelnen Optionen hintereinander eingegeben werden oder über eine *Graphical User Interface* (GUI) (siehe Abbildung 3.1).

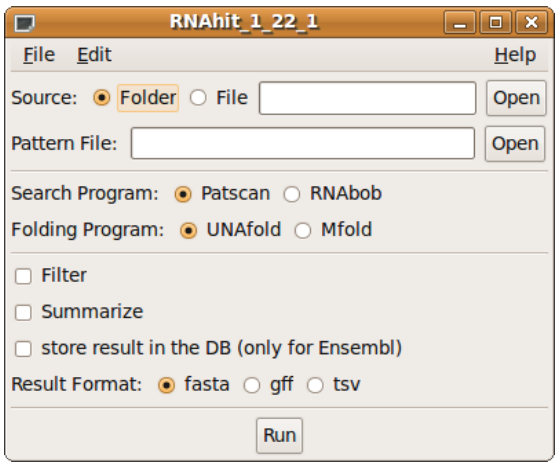
In der GUI gibt es drei Menüpunkte: *File, Edit* und *Help*.

Das *Help*-Menü zeigt eine kurze Zusammenfassung des Programms und gibt Hinweise zur weiteren Vorgehensweise (Abbildung 3.2).

Das *Edit*-Menü ermöglicht die Einstellung unterschiedlicher Optionen, die für RNAhit relevant sind (siehe Abbildung 3.3). Es kann die Anzahl zu verwendender Prozessoren eingegeben werden, die angibt, in wieviele *Threads* der Job aufgeteilt werden soll, also z. B. wieviele Suchen oder Faltungen parallel durchgeführt werden können. Standardmäßig wird mindestens ein Prozessor genutzt. Des Weiteren kann die prozentuale Menge des zu verwendenden Arbeitsspeichers angegeben werden. Falls diese Grenze erreicht wird, verharrt der Prozess und wartet bis Arbeitsspeicher aus anderen Prozessen wieder zurückgegeben wird, was einen Speicherüberlauf und damit Abbruch des Programmes verhindert.

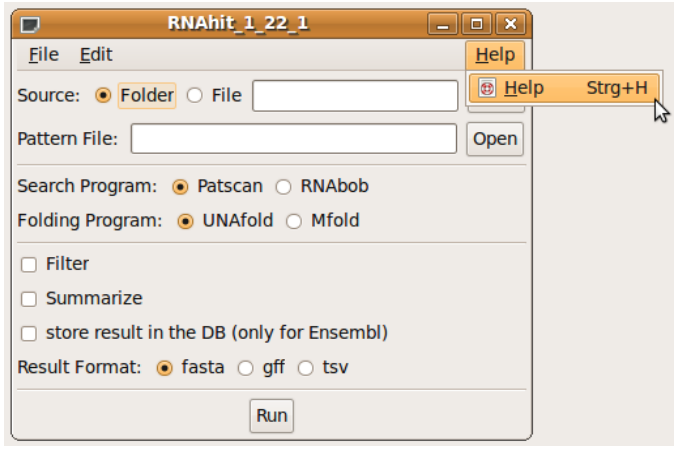


(a) Konsole

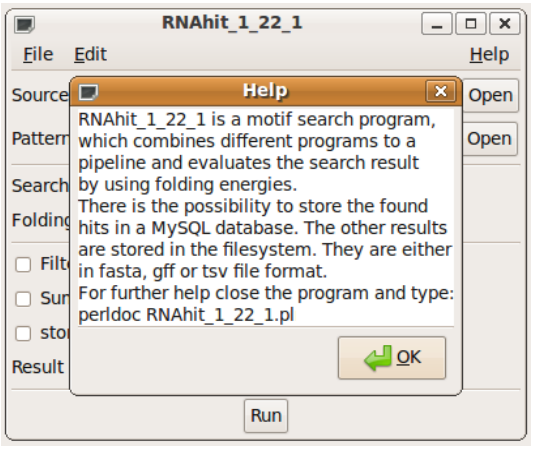


(b) RNAhit Benutzeroberfläche

Abbildung 3.1: RNAhit Aufruf
Der Programmaufruf geschieht über die Shell (a) oder über ein Programmfenster (b).



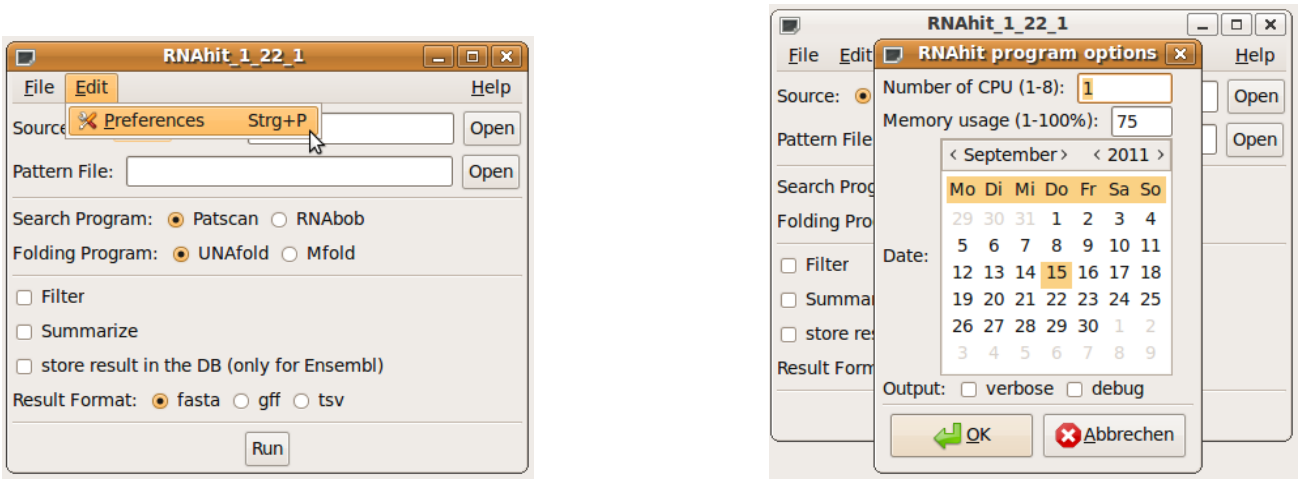
(a) RNAhit *Help* Aufruf



(b) RNAhit *Help* Auswahl

Abbildung 3.2: RNAhit *Help* Menü
Die Hilfe kann im Menü auf der rechten Seite aufgerufen werden (a). Daraufhin öffnet sich ein Hilfe Fenster (b).

Die Voreinstellung beträgt 75% des Arbeitsspeichers. Eine weitere Option ist die Einstellung eines Datums für eine Suche. Die Voreinstellung ist das aktuelle Datum des Programmstarts. Diese Option ermöglicht es eine abgebrochene Suche fortzuführen, die je nach Größe des Suchraums mehrere Monate andauern kann. Weitere Optionen sind *Checkboxes* für die kommentierte Ausgabe der einzelnen Schritte bzw. für einen Modus, der die Fehlersuche ermöglicht.



(a) RNAhit *Edit* Aufruf

(b) RNAhit *Edit* Auswahl

Abbildung 3.3: RNAhit *Edit* Menü
Das *Edit*-Menü befindet sich auf der linken Seite der Menüleiste (a). Es ermöglicht RNAhit spezifische Einstellungen (b).

Das *File*-Menü enthält drei Auswahlmöglichkeiten (siehe Abbildung 3.4): *Quit* zum Verlassen des Programmes, *Run* zum Starten der Suche und Download zum Herunterladen von neuen Genomen. Bei vorausgesetzter Internetverbindung können Genome von einem selbst vorgegebenen *File Transfer Protocol* (FTP) Server (Abbildung 3.5(a) und 3.5(b)) oder von Ensembl (Abbildung 3.5(c) und 3.5(d)) bzw. Ensemblgenomes (Abbildung 3.5(e) und 3.5(f)) heruntergeladen werden. Dabei besteht die Möglichkeit alle „*ALL LISTED*“ oder nur einen Organismus auszuwählen. Bei Ensemblgenomes muss zuvor eine Gruppe selektiert werden.

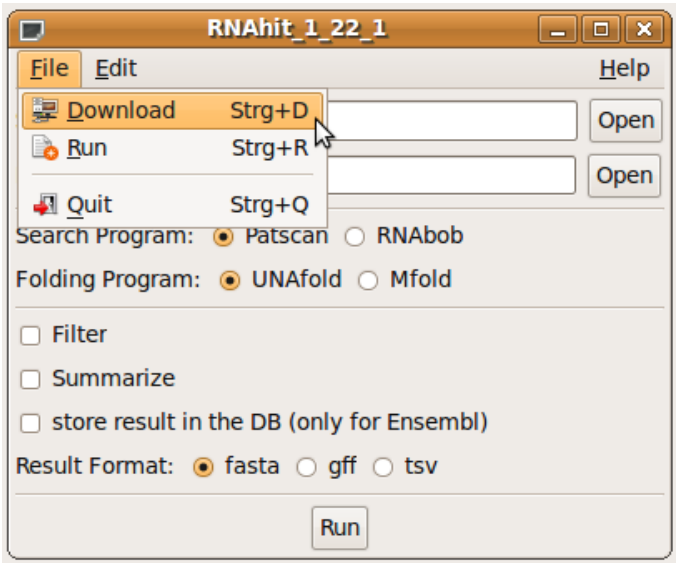
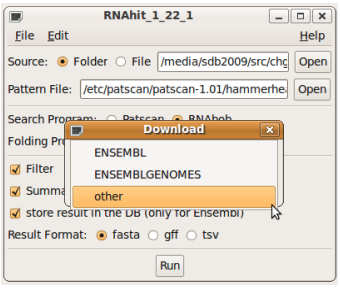
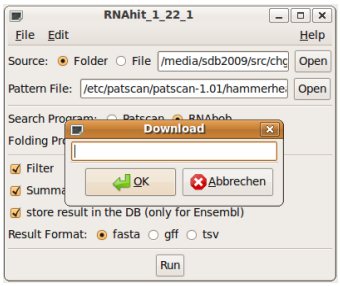


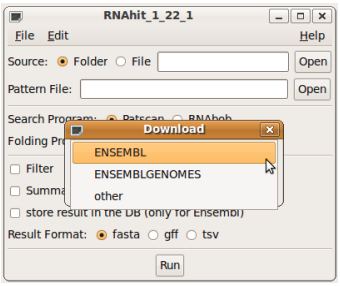
Abbildung 3.4: RNAhit *File* Menü
Das *File*-Menü bietet eine Download-Option zum Herunterladen neuer Genome, eine *Run*-Option zum Starten und eine *Quit*-Option zum Beenden des Programms.



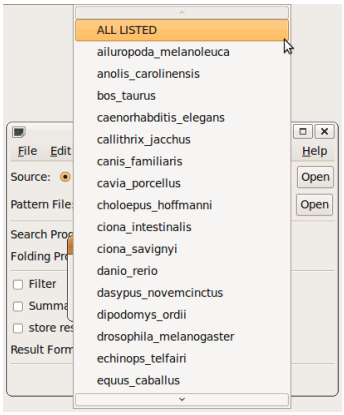
(a) RNAhit Download-Auswahl



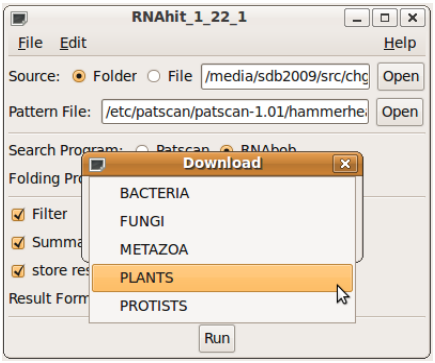
(b) RNAhit Download FTP Eingabe



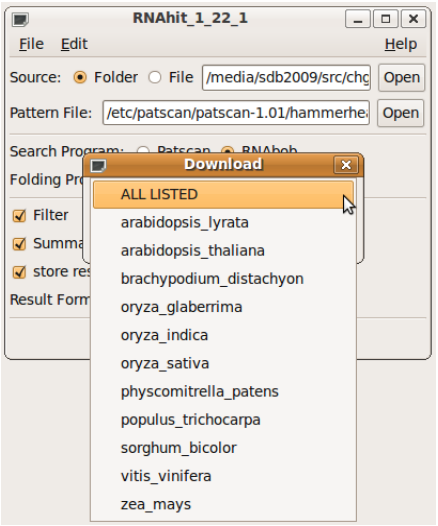
(c) RNAhit Download Ensembl



(d) RNAhit Download Ensembl Spezies



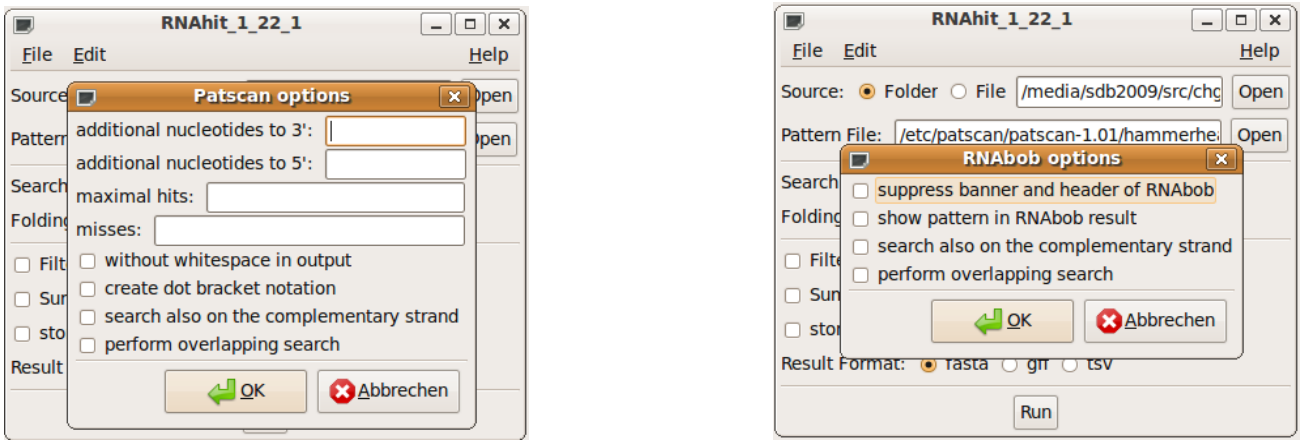
(e) RNAhit Download Ensemblgenomes Gruppe



(f) RNAhit Download Ensemblgenomes Spezies

Abbildung 3.5: RNAhit Download
RNAhit ermöglicht das Herunterladen neuer Genome aus verschiedenen Quellen.

Die Haupteingabemaske fragt alle weiteren benötigten Informationen ab, wie den Ordner bzw. die Datei, die durchsucht werden soll. Ein Ordner kann dabei einem oder mehreren Organismen entsprechen und eine Datei einem Genom oder einem Chromosom. Diese Ausgangsdaten sowie die für die Suche zu verwendende Beschreibung der Sekundärstruktur kann mit dem Pfad direkt eingegeben oder über „*Open*“ ausgewählt werden. Momentan werden zwei Suchprogramme angeboten deren Optionen durch Klicken auf die jeweiligen *Radiobutton* eingestellt werden können (siehe Abbildung 3.6(a) und 3.6(b)).



(a) RNAhit PatScan-Optionen

(b) RNAhit RNAbob-Optionen

Abbildung 3.6: RNAhit Suchprogramme
In RNAhit werden zwei *Pattern* Suchprogramme (PatScan (a), RNAbob (b)) mit unterschiedlichen Optionen eingebunden.

Bei den zwei eingebundenen Faltungsprogrammen ist keine optionale Änderung möglich, da jeweils ausschließlich die Standardeinstellungen genutzt werden (siehe Abschnitt 2.4.13). Die Filteroption ermöglicht eine spezifische Auswahl an Filtern, die je nach Motiv hinzugefügt oder entfernt werden können (siehe Abbildung 3.7).

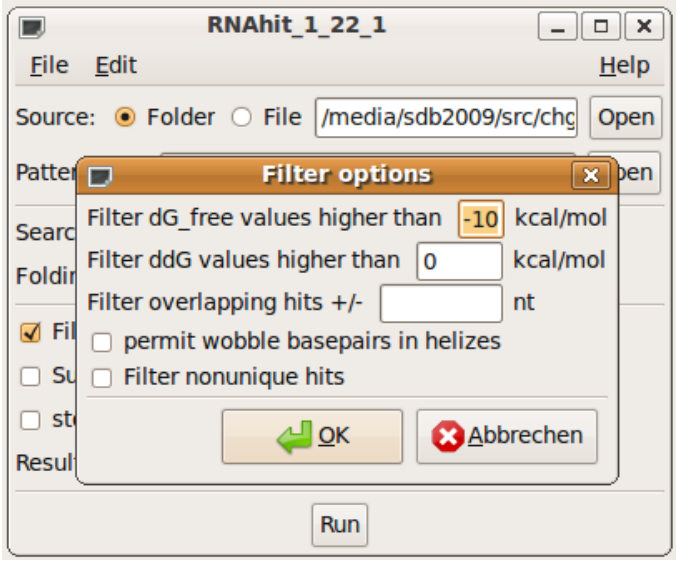


Abbildung 3.7: RNAhit Filter Optionen
Falsch positive Such- und Faltungsergebnisse können durch verschiedene Filtereinstellungen entfernt werden.

Wird die *Checkbox* „*Summarize*“ ausgewählt, wird zusätzlich, neben den Suchergebnissen pro Organismus, eine Zusammenfassung des Suchergebnisses in einer Datei erstellt. Die Ensembl Ergebnisse können ebenfalls zusätzlich in einer MySQL-Datenbank abgespeichert werden, wofür ein Passwort zur Authentifizierung abgefragt wird (Abbildung 3.8).

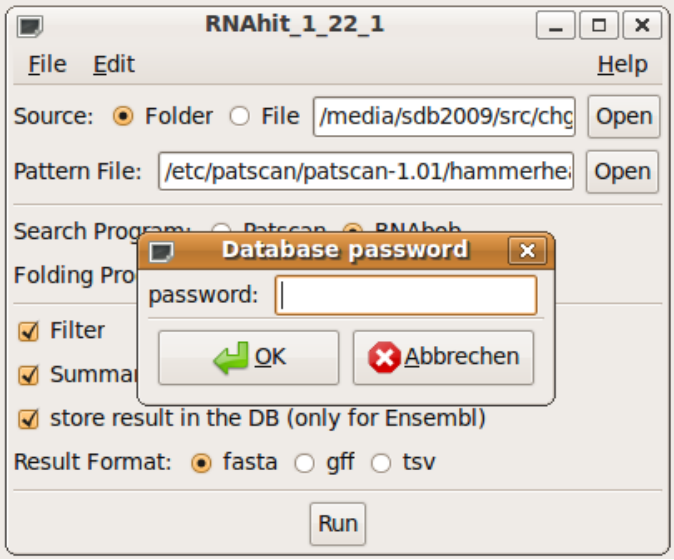


Abbildung 3.8: RNAhit Passwortabfrage
Die Datenbank erfordert aus Sicherheitsgründen für Schreibrechte zum Speichern gefundener Treffer ein Passwort.

Mit den letzten drei *Radiobutton* kann das Dateiformat der Ergebnisausgabe definiert und die Suche durch Drücken des „*Run*“-*Buttons* gestartet werden.

Nach dem Start werden zunächst verschiedene Perl-Module eingebunden, die während des Programmablaufes benötigt werden. Darunter befindet sich das eigens für dieses Programm geschriebene Modul Tools, welches häufig verwendete Methoden enthält, die ausgelagert wurden. Beispiele hierfür sind:

- der Wechsel von Verzeichnissen, das Verschieben von Dateien, die Ausgabe in eine Datei, das Packen und Entpacken von Dateien,
- Bestimmung des verfügbaren Speichers und der Zeit,
- statistische Auswertungen, wie das Ermitteln der Sequenzenanzahl und des GC-Gehaltes,
- das Einlesen der Punkt-Klammer-Struktur zur Bestimmung der *Constraints* und
- die Konvertierung von Dateiformaten (RNAbob zu PatScan, Fasta zu GFF, Fasta zu *Tab-separated Values* (TSV)).

Anschließend erfolgt die Definition und Initialisierung der verschiedenen Variablen, Listen und Hashes. In jedem Fall wird eine LOG-Datei erstellt, die das Datum, die Optionen und den Ablauf dokumentiert. In der Hauptroutine wird zunächst die Art des Programmaufrufes überprüft (siehe Abbildung 3.1). Je nach Aufruf erscheint ein Fenster mit einem Balken, der den Fortschritt des Programmes anzeigt (Abbildung 3.9).

Abbildung 3.10 zeigt eine Zusammenfassung des Programmablaufes.

Es besteht aus den Methoden

- checkInput,
- setOrganisms,
- getPatternResult,
- getFoldingResult,
- getFilterResult,
- getSummary,
- storeResult,
- cleanUp.

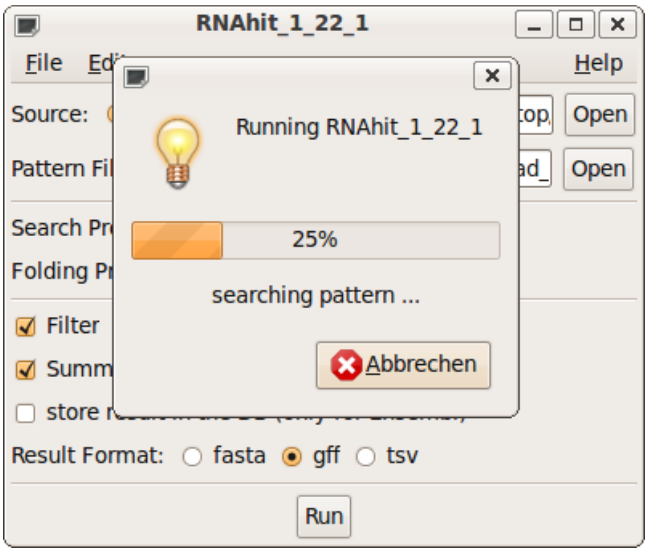


Abbildung 3.9: RNAhit Fortschrittsanzeige
RNAhit zeigt beim Aufruf über die GUI den Fortschritt der einzelnen Stufen an.

CheckInput überprüft die eingegebenen Optionen, stellt z. B. fest, ob es widersprüchliche Einstellungen gibt, kontrolliert die Umgebungsvariablen und setzt unterschiedliche Standardwerte. Diese Methode ist außerdem für die Erkennung vorheriger Suchen verantwortlich und für das Reparieren bzw. Entfernen beschädigter Dateien. Des Weiteren werden die zu durchsuchenden Dateien ermittelt.

SetOrganisms nutzt den Pfad der gegebenen Dateien und setzt den Ordnernamen als Organismus (siehe Abschnitt 2.4.17).

GetPatternResult überprüft ebenfalls, ob bereits Suchergebnisse vorhanden sind, die nicht erneut durchsucht werden müssen. Hinweise auf abgebrochene Suchen sind temporäre Dateien, die nicht entfernt wurden. Nach der Überprüfung wird je nach Anzahl zu verwendender Prozessoren das Suchprogramm mehrmals gestartet. Bevor ein neuer Prozess beginnt, wird die Größe des aktuell benötigten Arbeitsspeichers ermittelt. PatScan- und RNAbob-Optionen unterscheiden sich und müssen dementsprechend aufgerufen werden (siehe Abschnitt 2.4.12). Falls ein Motiv nicht in der Sequenz vorkommt, wird eine leere Ergebnisdatei erzeugt, um in späteren Analysen die Ausgangsmenge bestimmen zu können. Damit die darauffolgenden Schritte einheitlich bleiben, werden alle RNAbob-Ergebnisse in das PatScan-Format übertragen (Abschnitt 2.4.12). Diese Dateien besitzen den Präfix „p_“.

Anschließend folgt die Methode getFoldingResult. Zunächst werden alle gefundenen Suchergebnisse gesammelt. Je nach Anzahl der Prozessoren und verfügbarem Arbeitsspeicher wird die Methode getFolding aufgerufen. Die Idee hierbei besteht darin, für spätere Filterschritte die Sequenzen zu ermitteln, die wahrscheinlich die gleiche Funktion besitzen, da sie in die gleiche Struktur wie das gesuchte Motiv falten. Im Fall der Ribozyme ist dies die katalytische Aktivität. Dazu wird die freie Energie der gegebenen Sequenz bestimmt und diese mit der freien Energie der in das Motiv gezwungenen Sequenz verglichen. Dafür müssen Faltungsbeschränkungen (*Constraints*) formuliert werden (siehe Abschnitt 2.4.13). Die Methode getFoldingConstraints ermittelt, mit Hilfe der Punkt-Klammer-Struktur des Suchergebnisses und eigener Definitionen aus der Motivbeschreibungsdatei, die *Constraints*. *Constraints* für Motive, die zwei gleiche aufeinanderfolgende Module mit alternativen Längen besitzen - also zwei Einzelstränge oder zwei Helices hintereinander - können nicht erzeugt werden, da das erforderliche Aufteilen in Module aus der Punkt-Klammer-Struktur in solchen Fällen nicht ersichtlich ist. Falls es keine eigenen Definitionen gibt, werden alle Klammern (Helices) erzwungen und alle Punkte (Einzelstränge) verboten. Diese werden in eine temporäre Datei geschrieben und beim Programmaufruf eingebunden. Damit das Programm bei einer fehlgeschlagenen Faltung nicht abbricht, werden x, y, z für fehlende $\Delta G_{free}, \Delta G_{motif}$ und $\Delta\Delta G$ Werte eingetragen. Bei der Methode getFoldingResult erfolgt ebenfalls die Bestimmung der Nukleotidhäufigkeit, der Sequenzlänge und der Position der Spaltstelle.

Nachdem alle Such- und Faltungsergebnisse vorliegen („2_dG“-Dateien), erfolgt optional die Bewertung der Treffer durch die Filterung mittels gegebener Parameter. Dies geschieht in der Methode getFilterResult. Für jeden Filterschritt gibt es eine extra Datei. Sind alle Sequenzen der Ausgangsmenge herausgefiltert, weil ein bestimmtes Kriterium nicht erfüllt wird, fehlt die entsprechende Datei. Falls die jeweiligen Filter ausgewählt wurden, erfolgt die Filterung in der folgenden Reihenfolge. Zunächst werden alle Sequenzen, die über einem gegebenen ΔG_{free} Wert liegen entfernt.

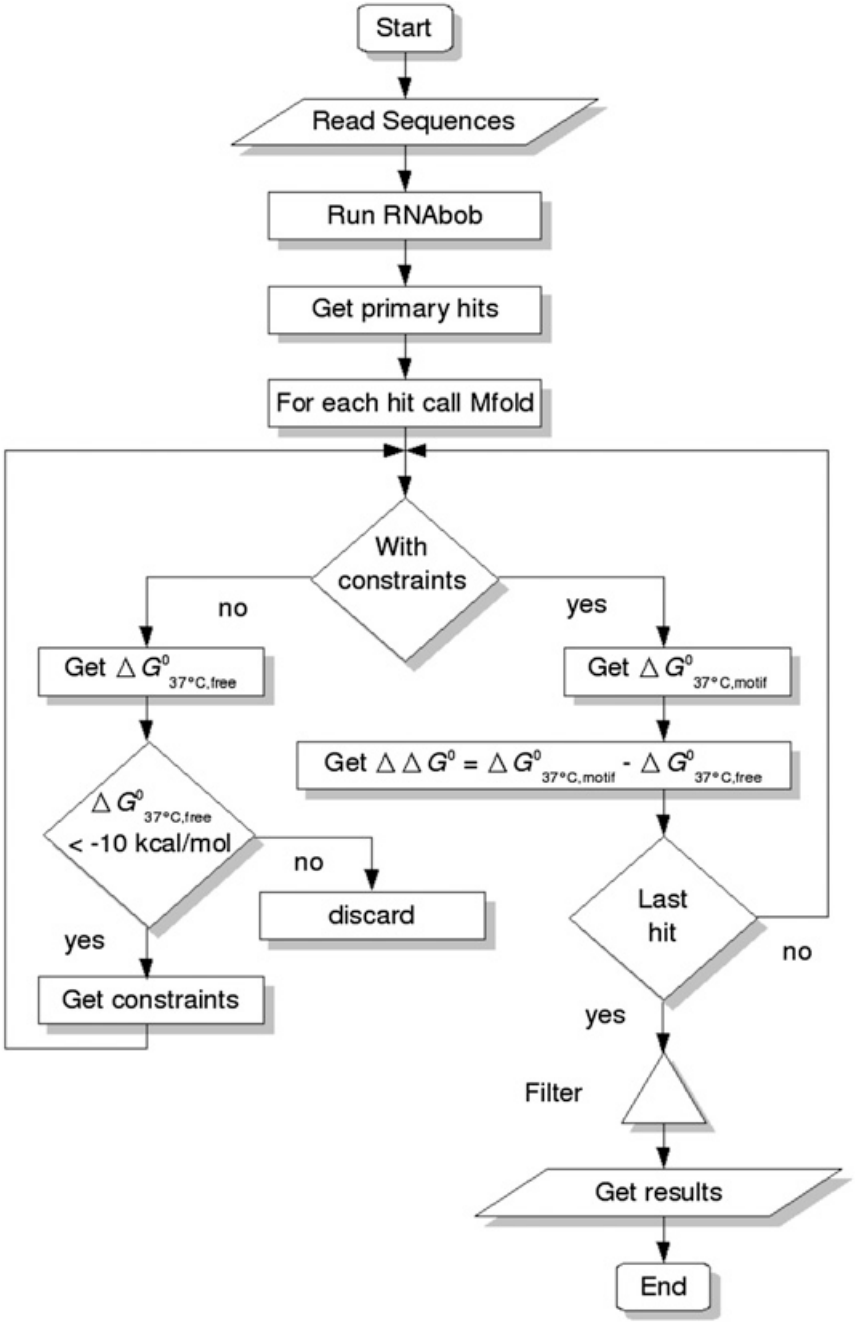


Abbildung 3.10: RNAhit *Pipeline*

Programmablaufplan von RNAhit [Seehafer et al., 2011]. Die Sequenzen werden aus verschiedenen Quellen eingelesen und mit einem Suchprogramm nach dem Motiv durchsucht. Die gefundenen Treffer werden dann frei und erzwungen gefalten und anschließend nach definierten Kriterien gefiltert (siehe Tabelle 2.3).

Das Ergebnis befindet sich in der „3_dGfree“-Datei. Diese wird für den nächsten Filterschritt verwendet, welcher alle Treffer filtert, die kleiner oder gleich der Differenz aus ΔG_{motif} und ΔG_{free}, dem $\Delta\Delta G$ Wert liegen („4_ddG“-Dateien). Falls überlappend gesucht wurde, werden alle alternativen Punkt-Klammer-Strukturen des Motivs derselben Region ausgegeben. Diese können innerhalb eines definiert Intervalls mit der Methode filterOverlappingHits gefiltert werden, wobei die Sequenz mit dem größeren $\Delta\Delta G$ Wert gelöscht wird. Das Resultat wird in der „nonoverlap_“-Datei ausgegeben. Ein weiterer Filter, der es ermöglicht nach einer Suche, bei der Wobble Basenpaare zunächst zugelassen wurden, diese nachträglich wieder zu entfernen, ist die Methode filterWobble. In dieser Methode werden alle Sequenzen entfernt, die mindestens ein Wobble Basenpaar in einer Helix besitzen („nonWobble_“-Dateien). Der letzte Filterschritt ist der Unique-Filter. Dieser Filter entfernt alle Sequenzen, die zu 100% identisch zu einer anderen Sequenz sind und zählt dabei wie häufig diese beobachtet wurde („unique_“-Dateien). Dabei sollte beachtet werden, dass dieser und alle anderen Filter sich auf eine Datei beziehen, welche je nach Datensatz einem kompletten Genom oder einem einzelnen Chromosom entspricht. Alle Filterschritte liegen ebenfalls unabhängig von RNAhit als einzelne Skripte vor und können im Nachhinein auf die Ergebnisse angewendet werden.

Die Methode getSummary fasst - falls ausgewählt - alle („unique_“-Dateien) in einer Fasta-Datei zusammen.

Angenommen die Ergebnisse sollen in der Datenbank gespeichert werden, erfolgt zunächst die Bestimmung der Datenbanknamen der Organismen, die bei Ensembl vorhanden sind. Wichtig ist die Datenbank auszuwählen, welche die richtige Release besitzt, da sonst Koordinaten- und Sequenzunterschiede vorkommen. Das Besondere der lokalen Datenbank ist, dass es sich um eine Multispezies-Datenbank handelt, da normalerweise für jeden Organismus eine Datenbank existiert. Die Organismen werden in der „meta“ Tabelle mit Verweisen („meta_coord“) zu ihren jeweiligen Koordinatensystemen („coord_system“) abgelegt. In der Methode storeResult wird zu Beginn überprüft, ob bereits Organismen in der Datenbank abgespeichert wurden, da sonst zwei unterschiedliche IDs zu einem Organismus vergeben werden und dies zu einem Datenbankfehler und damit zum Abbruch des Programmes führt. Anschließend wird die Datenbank des Organismus ausgewählt, eine Verbindung zur eigenen und zur Ensembl Datenbank hergestellt, notwendige Adaptoren geladen und über deren Funktionen die Treffer in die eigene Datenbank eingetragen, wobei jeder Organismus ein eigenes Koordinatensystem besitzt. Eine detaillierte Beschreibung der Koordinatensysteme und *Slices* befindet sich im Quellcode. Es existieren zwei Versionen der Datenbank. Anfangs wurden lediglich die gefilterten Treffer gespeichert, so wie in der aktuell geladenen Datenbank vom 06.08.2010 mit allen nach dem $\Delta G_{motif}/\Delta G_{free} = \Delta G_{ratio} > 0.6$ gefilterten Typ III *Hammerhead* Ribozymen. Inzwischen werden alle primären Such- und Faltungsergebnisse ungefiltert gespeichert. Ein Wechsel des Datenbank Backups ist im Abschnitt 2.4.11 beschrieben.

Der letzte Schritt des Programmes ist cleanUp - das Entfernen temporärer Dateien, die bei der Suche und Faltung entstanden sind.

countHits

Dieses Perl-Skript in Version 0.0.1 ist ein nützliches Werkzeug zum Ermitteln der Anzahl der Treffer und Ausgabe der Sequenzen. Dabei besteht die Möglichkeit, mit Hilfe der Levenshtein-Distanz ähnliche Sequenzen zu berücksichtigen und zusammenzufassen.

fasta2Tab

Wie der Programmname bereits besagt, kann mit diesem Skript in Version 0.0.5 eine Fasta-Datei in eine *Tab-separated Values* (TSV)-Datei umgewandelt werden. Vorausgesetzt wird, dass ein Fasta-Eintrag aus 3 Zeilen besteht und die Punkt-Klammer-Struktur enthalten ist.

fasta2gff

Dieses Skript (Version 0.0.1) konvertiert ähnlich zum Vorherigen eine Fasta-Datei in eine GFF-Datei. Punkt-Klammer-Strukturen in der Fasta-Datei werden vorausgesetzt. Beide Skripte sind in dem Modul Tools und somit in RNAhit enthalten.

getSpecies

Das Perl-Skript getSpecies_0_0_1.pl dient dem automatischen Herunterladen großer Datenmengen aus verschiedenen Quellen (siehe Abschnitt 2.1). Dazu kann ein bestimmter Host mit der Option -H übergeben werden. Standardeinstellung ist die Verwendung des Hosts „ftp.ensembl.org“. Des Weiteren wird ein Download-Pfad gesetzt. Nach dem Verbindungsaufbau erfolgt der Wechsel in das Verzeichnis des Organismus, wo gezielt nach gepackten Dateien gesucht wird. Anschließend wird im Download-Pfad ein Ordner erstellt, der den Speziesnamen trägt, falls dieser noch nicht existiert. Danach werden die Dateien in diesen Ordner übertragen. Je nach Host muss die Verzeichnisstruktur des Servers berücksichtigt werden. Das Skript wurde zum Teil für Ensembl und Ensemblgenomes in RNAhit aufgenommen.

createFilesFromGbk

Ein Teil der Sequenzdaten liegt ausschließlich im GenBank-Format vor. Diese müssen für die Suchprogramme in das Fasta-Format umgewandelt werden. Dazu wird zunächst ermittelt, welche Organismen bereits existieren. Anschließend werden aus der gegebenen gbk Datei die ID, die Sequenz und der Organismus bestimmt. Falls noch kein Ordner mit dem Speziesnamen existiert, wird dieser erstellt und darin die neue Fasta-Datei angelegt.

ROCscript

Dieses Skript dient der automatischen Parameterwahl für RNAhit. Dazu wird eine Liste von „richtig positiven“ (TP), also Sequenzen von bekannten katalytisch aktiven *Hammerhead* Ribozymen aus subviral DB [Rocheleau & Pelchat, 2006] und Anne Kalweits Experimenten, erstellt. Eine Liste von „falsch positiven“ Sequenzen, die das HHRz-Motiv enthalten, aber nicht aktiv sind, wurde ebenfalls von Anne Kalweit experimentell getestet und zur Verfügung gestellt. In beiden Listen (44 FP und 173 TP) wird in einer Schleife jeweils mit steigendem ΔG_{free} Wert, $\Delta\Delta G$ Wert und Überlapp-Wert sowie mit und ohne Unique-Filter, Wobble-Filter, Mfold / UNAfold und PatScan / RNAbob gesucht. Das jeweilige Ergebnis setzt sich aus zwei Zeilen zusammen, bestehend aus der Anzahl der TP (Zeile 1) und der Anzahl der FP (Zeile 2). Die Ergebnisse werden zum erstellen der ROC-Kurve (Abbildung 3.22) genutzt.

countFreqOfPos

Dieses Skript (Version 1.2.0) zählt mit Hilfe einer gegebenen Motivbeschreibung in einem Suchergebnis alle beobachteten Nukleotide bestimmter Positionen z. B. 3 und 8 im HHRz gemäß [Hertel et al., 1992] (siehe Abbildung 3.24).

filterC3G8

Neben den anderen Filterprogrammen sucht dieses Skript in Version 0.0.1 mit Hilfe einer gegebenen Motivbeschreibung speziell im HHRz nach Sequenzen, die an den Positionen 3 und 8 ein CG Basenpaar besitzen. Dafür wird eine gegebene Punkt-Klammer-Struktur vorausgesetzt.

splitFile

Große Dateien können mit diesem Skript (Version 0.0.1) in kleine Dateien unterteilt werden. Dazu wird die Anzahl der Zeilen ermittelt und daraus die Partitionsgröße berechnet, je nachdem wieviel Teile erstellt werden sollen. Es dient z. B. der Aufteilung von Suchergebnissen, um diese im Nachhinein parallel falten zu können.

getNCBIhits

Zur weiteren Analyse gefundener Treffer werden 5’ und 3’ zusätzliche Sequenzinformationen benötigt. Bei Ensembl besteht die Möglichkeit diese Informationen schnell mit wenigen Befehlen über die EnsemblAPI zu ermitteln. Bei allen anderen Quellen ist dies jedoch nicht möglich. Deshalb wurde das Skript getNCBIhits erstellt.

Es liest zunächst das Ergebnis im TSV-Format ein und ermittelt den Namen und Systempfad des Organismus. Anschließend wird mit PatScan erneut nach der Sequenz im Genom gesucht. PatScan bietet dabei die Option zusätzliche Nukleotide neben dem Treffer auszugeben. Die erhaltene, erweiterte Sequenz wird anschließend in eine neue Datei geschrieben.

getPosition

Ein ähnliches Programm zum Ermitteln von Sequenzdaten, jedoch mit einem anderen Ansatz, ist das Skript getPosition.pl. Mittels einer Datei mit einer oder mehreren Sequenzen, einem Chromosomennamen sowie einer Start- und Endposition sucht es die entsprechende Sequenz heraus und gibt sie aus. Dies ist ebenfalls für den Minusstrang möglich, wobei die Startposition größer als die Endposition ist und dann das *Reverse Complement* gebildet und ausgegeben wird.

getReverseComplement

Falls für eine Fasta-Datei mit einer oder mehreren Sequenzen das *Reverse Complement* gebildet werden soll, ist dieses Skript nützlich. Das Ergebnis wird in einer neuen Datei ausgegeben.

getEST

Das Skript getEST.pl dient ebenfalls der Auswertung der „nicht-Ensembl“ Treffer. Es überprüft, ob die gegebene Sequenz in der EST Datenbank vorkommt, was bedeuten würde, dass die Sequenz und damit das Motiv exprimiert wird. Ist das der Fall, so müsste für die *Hammerhead* Ribozyme die Sequenz gespalten vorliegen, da es sich sonst um kein HHRz handelt. Dementsprechend wird nach EST gesucht, die entweder mit der Spaltstelle enden oder beginnen. Dazu wird das Ergebnis zunächst im TSV-Format eingelesen und jede Sequenz künstlich in zwei Spaltprodukte gespalten. Anschließend wird für jedes Spaltprodukt eine BLASTN Suche gegen die EST Datenbank durchgeführt und der Start bzw. das Ende der ESTs mit der Sequenz verglichen.

getSimilarity

Dieses Perl-Skript in Version 0.2.0. erstellt eine Levenshtein-Distanzmatrix aus den im Fasta-Format gegebenen Sequenzen zur anschließenden Visualisierung der Sequenzähnlichkeit, z. B. mit Hilfe einer *Heatmap* (siehe Abbildung 3.21).

getL1L2

Verschiedene getestete MSA Algorithmen, wie z. B. ClustalW [Thompson et al., 1994], Muscle [Edgar, 2004] und T-Coffee [Notredame et al., 2000], welche strukturelle Informationen der RNA nicht einbeziehen, waren nicht in der Lage, das konservierte katalytische Zentrum des HHRz zu erkennen, obwohl dieses in jeder Sequenz enthalten ist. In dem MSA wurden in den konservierten Regionen mehrere *Gaps* eingefügt. Dies liegt - u. a. aufgrund variabler Positionen - an den großen Längenunterschieden der Motive. Da ein gutes MSA Voraussetzung zum Erstellen eines phylogenetischen Baumes ist, wurde das Skript getL1L2.pl geschrieben. Es nutzt die Punkt-Klammer-Struktur der Sequenzen, um Helix- und Einzelstrangregionen auszugeben. Dadurch können Teilsequenzen der Motive erstellt, einzeln aligniert und anschließend wieder zusammengefügt werden.

getFrequencyOfHitInGenome

Dieses Skript in Version 0.0.5 ermittelt, wie häufig ein gefundener Treffer in einem Genom vorkommt. Dazu werden zunächst alle Treffer eingelesen und ein Datensatz mit einzigartig vorkommenden Sequenzen erzeugt. Optional können diese anschließend mit BLASTN gesucht und gezählt werden, wobei eine für das Genom vorformatierte Datenbank vorausgesetzt wird (siehe Abschnitt 2.4.12).

getRTEannotation

Dieses Skript, welches der Annotation retrotransposabler Elemente dient, wurde bei der Auswertung der HDV-Ribozym Treffer in *Dictyostelium discoideum* verwendet, da Treffer in Retrotransposons beobachtet wurden. Es ermittelt zu einer gegebenen Gen ID den dazugehörigen Proteinnamen.

getObservedMotifProbability

Wie in Abschnitt 2.2 bereits erläutert kann die Berechnung von Motivwahrscheinlichkeiten aufgrund variabler Helix- und *Loop*-Längen sehr komplex werden. Um Wahrscheinlichkeiten dennoch abschätzen zu können, wurde das Skript getObservedMotifProbability Version 0.3.0 erstellt. Zu Beginn wird das Motiv und die Anzahl der Such-Wiederholungen pro Sequenzlänge abgefragt. Es gibt zwei Möglichkeiten das Programm aufzurufen. Zum einen kann eine Datei mit Genomgrößen übergeben werden, wobei die Größe als Vorlage der Zufallssequenzlänge dient oder es wird für die Länge eine Startgröße übergeben, die nach den gegebenen Wiederholungen beim nächsten Durchlauf verzehnfacht wird, solange bis eine obere Grenze erreicht wird. Die maximale Zufallssequenzlänge liegt, aufgrund gegebener Hardware-Limitierung, bei 2 Milliarden Nukleotiden. In den mit dem Skript gen_nucleic_stefan_graef.pl (siehe Abschnitt 2.4.14) erstellten Sequenzen wurde anschließend nach dem Motiv gesucht und gezählt, wie häufig es beobachtet werden konnte. Optional können aus dem Ergebnis überlappende Treffer herausgefiltert werden.

getSupportingEvidence

Dieses Skript in Version 0.1.3 baut auf ein Skript von Stefan Gräf auf, welches sich mit der Ensembl-Schnittstelle verbindet und Annotationen zu gegebenen Treffern aus der Datenbank heraussucht. Dies ist aufgrund der Schnittstelle nur für Ensembl und Ensemblgenomes Spezies möglich. Der erste Schritt ist das Einlesen der Treffer im TSV-Format. Daraus wird die Release der jeweiligen Datenbank ermittelt. Je nach Spezies wird anschließend eine Verbindung zu Ensembl bzw. zum European Bioinformatics Institute (EBI) aufgebaut. Danach werden über den Datenbank- und *Slice Adaptor* alle Annotationen der Region des Treffers abgefragt und in einer Datei ausgegeben.

3.1.2 R

ROCRscript

Dieses R-Skript liest das Ergebnis der gefundenen Scores (Anzahl TP, FP) ein und erstellt unter Verwendung der Bibliothek ROCR [Sing et al., 2005] eine ROC-Kurve (Abbildung 3.22).

HHRzI

Bei diesem Skript werden zunächst zwei Dateien eingelesen. Zum einen die ungefilterte Anzahl gefundener HHRz-Motive Typ I in verschiedenen Organismen und zum anderen die ungefilterte Anzahl gefundener HHRz-Motive Typ I in Zufallssequenzen entsprechend der Genomgröße der jeweiligen Organismen. Die Anzahl und Genomgröße wird, wie in Abbildung 3.25 zu sehen, logarithmisch gegeneinander aufgetragen. Die Trefferanzahl der Zufallssequenzen wird anschließend in einem Mittelwert zusammengefasst und aus den Mittelwerten und den Genomgrößen ein lineares Modell gebaut. Mit diesem ist es anschließend möglich, Werte des Konfidenzintervalls vorherzusagen und einzuzeichnen. In einem letzten Schritt werden zusätzlich Histogramme der Genomgrößen und Trefferanzahlen hinzugefügt.

3.1.3 Shell

Viele der erstellten Perl-Skripte sind im Verlauf der Arbeit durch Shell-Skripte ersetzt worden, welche mit wenigen Befehlen das gleiche Ergebnis erzielen. Teilweise ist Perl in den Skripten integriert. Nachfolgend wird ein Auszug der vorhandenen Skripte vorgestellt.

HHRzIII_03122010

Dieses Skript dient der Zusammenfassung von Suchergebnissen und dabei dem Ermitteln der Gesamttreffermenge. Dazu wird nach Dateien gesucht, die das Datum „3Dec2010“ und das Motiv „typIII“ im Dateinamen enthalten. Anschließend werden die Rohdaten herausgesucht und der Inhalt dieser Dateien in einer Fasta-Datei zusammengefasst.

HHRzI_2dG_07012011

Im Gegensatz zum vorherigen Skript sucht dieses nach bereits gefalteten Suchergebnissen, welche an dem „2_dG_“ Präfix erkennbar sind. Außerdem wird für spätere Stastiken die Sequenzanzahl pro Datei, die Anzahl der Treffer pro Organismus und die Genomgröße ermittelt.

getPrimaryData

Dieses Skript in Version 0.1.4 und das Skript getFilteredData.sh sind Weiterentwicklungen der vorherigen zwei Skripte und dienen der Auswertung der Daten. Es wird mit zwei Parametern (Datum und Motiv) aufgerufen. Anschließend wird eine Liste von Dateien erstellt, die das Datum und das zu suchende Motiv im Dateinamen enthalten. Diese Liste entspricht den primären Suchergebnissen. Danach wird der Benutzer dazu aufgefordert auszuwählen, ob die Dateien in einem mit dem Datum erstellten Ordner auf den Desktop kopiert werden sollen oder der Inhalt aller Dateien in einer Fasta-Datei auf dem Desktop abgespeichert werden soll. Die Dateiliste wird außerdem dazu verwendet, Statistiken über die Anzahl der Treffer in Viren, Archaeen, Bakterien und Eukaryonten zu erstellen.

getFilteredData

Das Skript getFilteredData in Version 0.1.3 funktioniert nach einem ähnlichen Prinzip wie getPrimaryData mit dem Unterschied, dass drei Parameter beim Aufruf übergeben werden und somit neben Datum und Motiv ein Filter ausgewählt werden kann. Auf diese Weise ist es möglich, z. B. alle „unique“ Dateien eines Datums auf den Desktop zu kopieren oder zu sammeln. Dabei werden ebenfalls Statistiken erstellt.

getTotal

Dieses Shell-Skript sucht nach allen vorhandenen gepackten Genomdaten und erstellt Statistiken über die Gesamtdatenmenge, die sich aktuell auf dem Server befindet.

Bowtie2_SamTools

Bowtie2_SamTools in Version 0.0.7 zeigt beispielhaft den Aufbau eines sbatch Shell-Skripts. Auf 4 Knoten des Clusters Kalkyl werden in einem Zeitraum von 5 Tagen die *Reads* aus dem Sequenzierungsergebnis gegen das *Dictyostelium discoideum* Genom mit einer sehr sensitiven Parametereinstellung aligniert. Die erhaltenen SAM-Dateien werden anschließend in komprimierte BAM-Dateien umgewandelt, nach der *Alignment*-Position sortiert und indexiert. Am Ende erfolgt der Aufruf des nächsten Skriptes, so dass eine *Pipeline* entsteht.

Cuff_dicty

Ein sbatch Skript, das im Anschluss an das *Mapping* aufgerufen werden kann, ist z. B. Cuff_dicty Version 0.0.14. Es besteht aus cufflinks, cuffcompare sowie cuffdiff und dient der Bestimmung signifikanter Unterschiede zwischen den *Read*-Häufigkeiten.

HTseq_dicty

Das sbatch Skript HTseq_dicty Version 0.0.18 zählt mit Hilfe der von dictybase gegebenen Annotationen, wieviel *Reads* mit einer bestimmten Region überlappen.

3.1.4 MySQL

deleteDB

Mit Hilfe der Datei deleteDB.sql kann der Inhalt der entsprechenden Tabellen des Ensembl-Schemas gelöscht werden. Das Entfernen der Spezies aus der „meta“-Tabelle muss jedoch manuell ausgeführt werden (siehe Abschnitt 2.4.11).

3.2 Suchergebnisse

Die Ergebnisse befinden sich auf dem Server (IP-Adresse: 141.50.190.146) in einer MySQL-Datenbank und im Dateisystem unter

`/media/sdb2009/src/`

(siehe Abschnitt 2.4.17). Darunter zählen die Such- und Filterergebnisse der einzelnen Motivsuchen, die zum Teil in verschiedenen Dateiformaten, wie Fasta, TSV oder GFF vorliegen. Ein Beispiel für eine bei Ensembl hochgeladene GFF-Datei und der Darstellung gefundener Treffer innerhalb einer annotierten Umgebung ist in Abbildung 3.11 zu sehen.

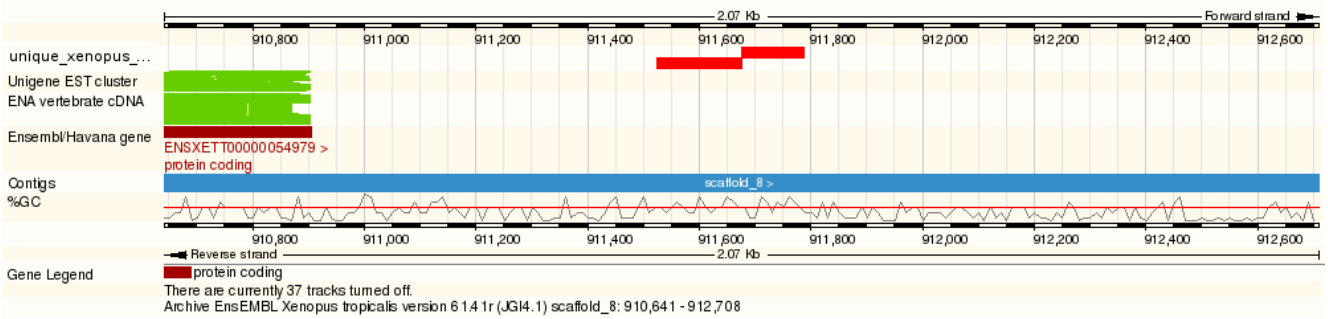


Abbildung 3.11: Ensembl *Xenopus tropicalis* (JGI4.1)
Die roten Balken zeigen zwei sich überlappende, mögliche *Hammerhead* Ribozyme Typ III aus *Xenopus tropicalis*. Diese befinden sich im intergenischen Bereich auf dem Plusstrang von scaffold_8.

Statt der GFF-Datei kann alternativ bei Ensembl unter „*Manage your data* > *Attach DAS*“ ein Distributed Annotation System-Server angegeben werden. Dies wurde jedoch aufgrund der Firewall Einstellungen der Universität Kassel nicht getestet. In Abbildung 3.12 ist ein den Daten angepasster ProServer zu sehen, erreichbar nach Aktivierung (siehe Abschnitt 2.4.16) unter

http://localhost:9008/das/sources/

Die dargestellte Tabelle ist das Ergebnis einer Anfrage an den ProServer Daemon, welcher eine DAS-*Query* an den Source Adaptor (MotifAdaptor.pm) formuliert. Daraufhin wird über den Transport Adaptor (MotifTransport.pm) eine SQL Abfrage an die Datenbank gesendet und die erhaltenen Daten werden in eine XML-Datei umgewandelt. Diese wird anschließend mit Hilfe von XSL-*StyleSheets* auf dem Bildschirm des Benutzers ausgegeben. Wählt der Benutzer durch das Betätigen eines Links einen Organismus aus, z. B. *Xenopus tropicalis*, werden die gleichen Schritte (siehe Abschnitt 2.4.16) durchgeführt und, wie in Abbildung 3.13, auf dem Bildschirm ausgegeben. Die Abbildung zeigt neben den Positionsangaben die berechnete freie Energie $\Delta G_{ratio} = \Delta G_{motif} / \Delta G_{free}$.

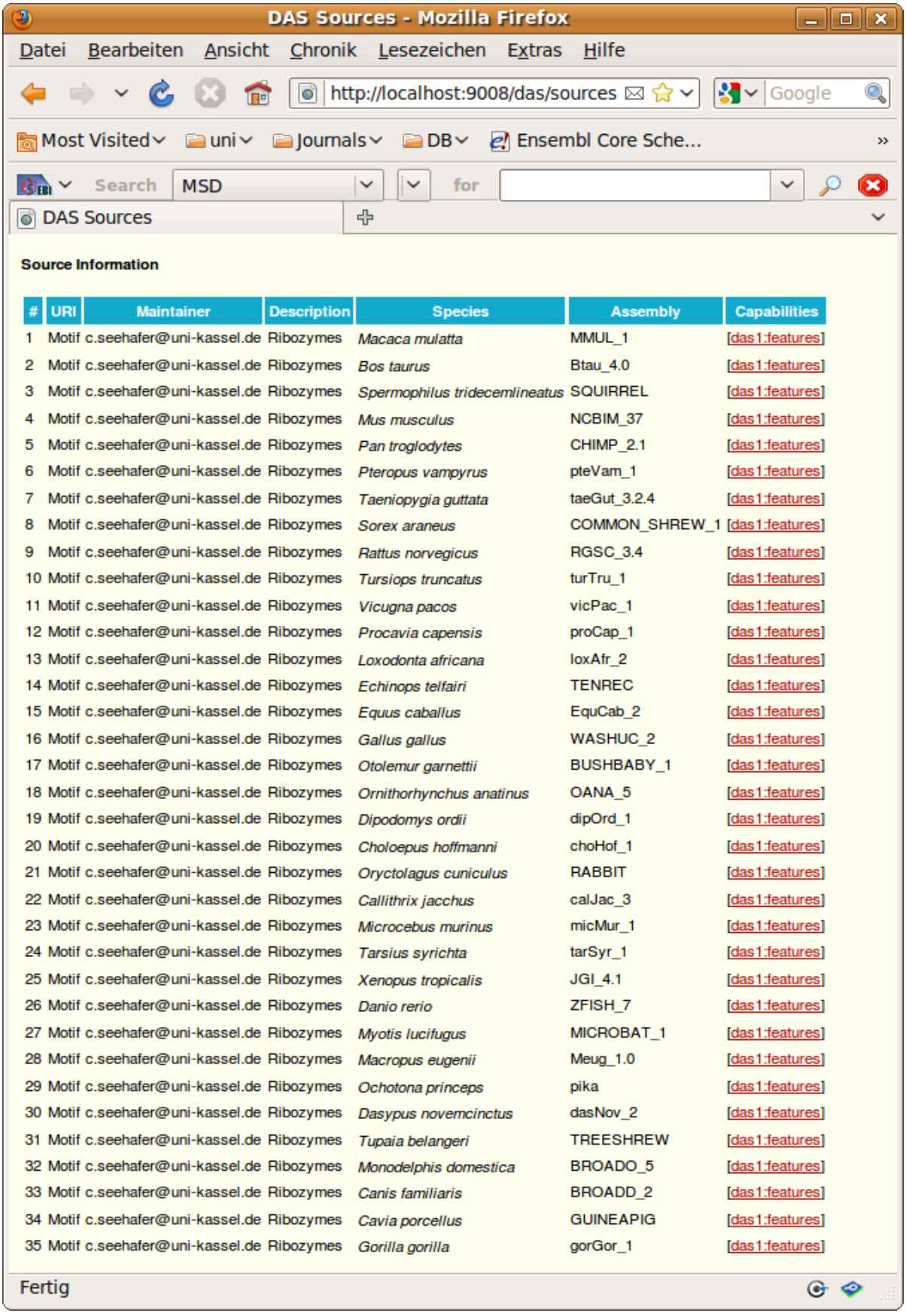
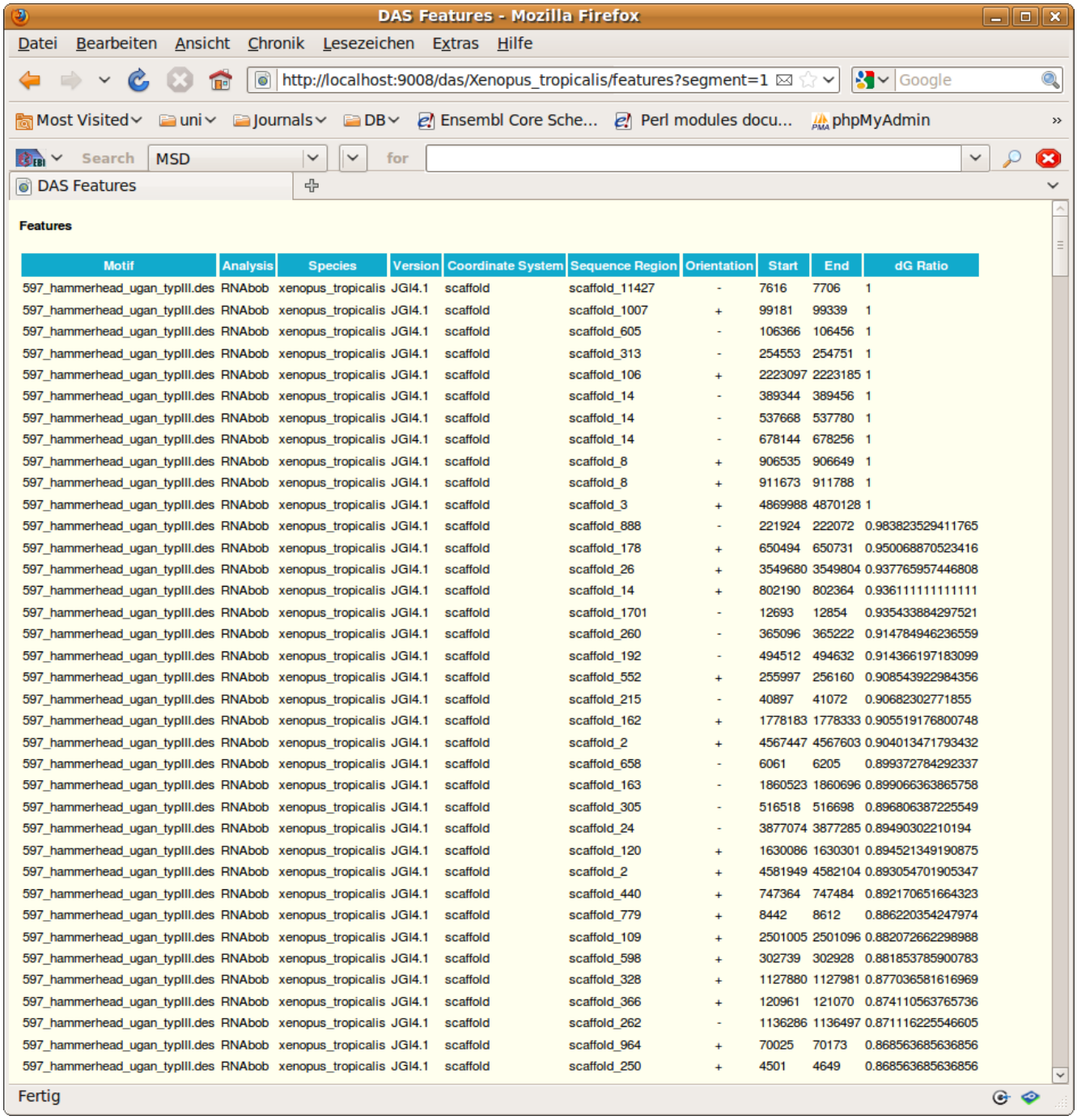


Abbildung 3.12: ProServer: Speziesüberblick

Dieser Ausschnitt des ProServers zeigt diverse Spezies mit ihren jeweiligen *Assembly*-Versionen, in denen die verschiedenen Motive gefunden wurden und die durch Betätigung des Links aus der Datenbank abgefragt werden können.



Motif	Analysis	Species	Version	Coordinate System	Sequence Region	Orientation	Start	End	dG Ratio
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_11427	-	7616	7706	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_1007	+	99181	99339	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_605	-	106366	106456	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_313	-	254553	254751	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_106	+	2223097	2223185	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_14	-	389344	389456	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_14	-	537668	537780	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_14	-	678144	678256	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_8	+	906535	906649	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_8	+	911673	911788	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_3	+	4869988	4870128	1
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_888	-	221924	222072	0.983823529411765
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_178	+	650494	650731	0.950068870523416
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_26	+	3549680	3549804	0.937765957446808
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_14	+	802190	802364	0.936111111111111
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_1701	-	12693	12854	0.935433884297521
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_260	-	365096	365222	0.914784946236559
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_192	-	494512	494632	0.914366197183099
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_552	+	255997	256160	0.908543922984356
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_215	-	40897	41072	0.90682302771855
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_162	+	1778183	1778333	0.905519176800748
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_2	+	4567447	4567603	0.904013471793432
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_658	-	6061	6205	0.899372784292337
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_163	-	1860523	1860696	0.899066363865758
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_305	-	516518	516698	0.896806387225549
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_24	-	3877074	3877285	0.89490302210194
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_120	+	1630086	1630301	0.894521349190875
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_2	+	4581949	4582104	0.893054701905347
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_440	+	747364	747484	0.892170651664323
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_779	+	8442	8612	0.886220354247974
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_109	+	2501005	2501096	0.882072662298988
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_598	+	302739	302928	0.881853785900783
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_328	+	1127880	1127981	0.877036581616969
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_366	+	120961	121070	0.874110563765736
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_262	-	1136286	1136497	0.871116225546605
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_964	+	70025	70173	0.868563685636856
597_hammerhead_ugan_typIII.des	RNAbob	xenopus_tropicalis	JGI4.1	scaffold	scaffold_250	+	4501	4649	0.868563685636856

Abbildung 3.13: ProServer: *Xenopus tropicalis*

Nach Auswahl des Organismus *Xenopus tropicalis* werden die gefundenen HHRz III mit Positionsangaben und der berechneten freien Energie $\Delta G_{motif}/\Delta G_{free} = \Delta G_{ratio}$ aus der Datenbank abgefragt und ausgegeben.

Zusätzlich sind die gefilterten Ergebnisse je nach Suchdatum im Anhang der Arbeit abgelegt.

Es wurde nach *Hairpin*, *Hammerhead*, HDV und VS ähnlichen Ribozymen gesucht. Die Suchergebnisse der verschiedenen Daten (siehe Tabelle 2.1) werden in den nachfolgenden Abschnitten zusammengefasst.

3.2.1 Hairpin Ribozyme

Tabelle 3.1 zeigt eine Übersicht der Suchergebnisse nach *Hairpin* Ribozymen. Die Suchparameter und Deskriptoren befinden sich im Abschnitt 2.4.17.

[5]

Tabelle 3.1: *Hairpin* Ribozym Ergebnis

Datum	Suche	Faltung	Spezies	Primär	Filter
26. Okt. 2009	PatScan	keine	199V 5A 595B 112E	0V 0A 0B 1 (1E)	-
11. Jan. 2010	RNAbob	keine	1371V 0A 615B 172E	0V 0A 0B 1 (1E)	-
14. Jan. 2010	RNAbob	Mfold	1373V 0A 615B 172E	0V 0A 0B 16 (13E)	0V 0A 0B 1 (1E)
17. Mär. 2010	RNAbob	Mfold	1373V 0A 617B 173E	9 (1V) 0A 42 (12B) 1109 (99E)	0V 0A 2 (2B) 74 (42E)

[5] Übersicht der Suchen nach *Hairpin* Ribozymen mit den jeweiligen Such- und Faltungsprogrammen. Die grauen Balken entsprechen in der „Spezies“ Spalte der Gesamtanzahl durchsuchter Spezies, in der „Primär“ Spalte der Anzahl primärer Treffer des Suchprogrammes pro Gruppe sowie in der „Filter“ Spalte der Anzahl gefilterter Treffer pro Gruppe, jeweils mit der Gruppenanzahl der Spezies in Klammern. Die Spezies werden in Viren (V), Archaeen (A), Bakterien (B) und Eukaryonten (E) unterteilt. Viroide und Satelliten RNA wurden in dieser Übersicht den Viren zugeordnet.

Die erste genomweite Suche nach einem Teilmotiv des *Hairpin* Ribozyms (Abbildung 2.11(a)), erfolgte mit PatScan und ohne weitere Filterungen. Die Suche ergab einen einzigen Treffer auf Chromosom 5 (*Assembly* BROADO5) an Position 100028160 - 100028196 im Opossum (*Monodelphis domestica*). Eine nähere Untersuchung dieser Region zeigte jedoch, dass weitere Teile des Motivs, wie die Spaltstelle, fehlen. RNAbob als alternatives Suchprogramm, ermittelte in einer erweiterten Suchmenge denselben Treffer (11.01.2010). Wird das Motiv im *Loop* allgemeiner formuliert, wie in Abbildung 2.11(c), werden 16 mögliche *Hairpin* Ribozyme in 13 Spezies gefunden, von denen nach der Analyse und Filterung 1 Treffer in Orang-Utan (*Pongo pygmaeus*) *Assembly* PPYG2, Chromosom 1: 93565412 bis 93565448 auf dem Minusstrang verbleibt. Eine nähere Überprüfung zeigt jedoch, dass auch bei diesem Treffer weitere Teile des Motivs fehlen (siehe Abbildung 3.14). Eine dritte Verallgemeinerung ist die Erweiterung des *Loops* auf bis zu 100 nt. Diese Suche ergab zwei Treffer in zwei Bakterien und 74 Treffer in 42 Eukaryonten, die der *Pattern*-Beschreibung und Faltung des Teilmotivs entsprechen.

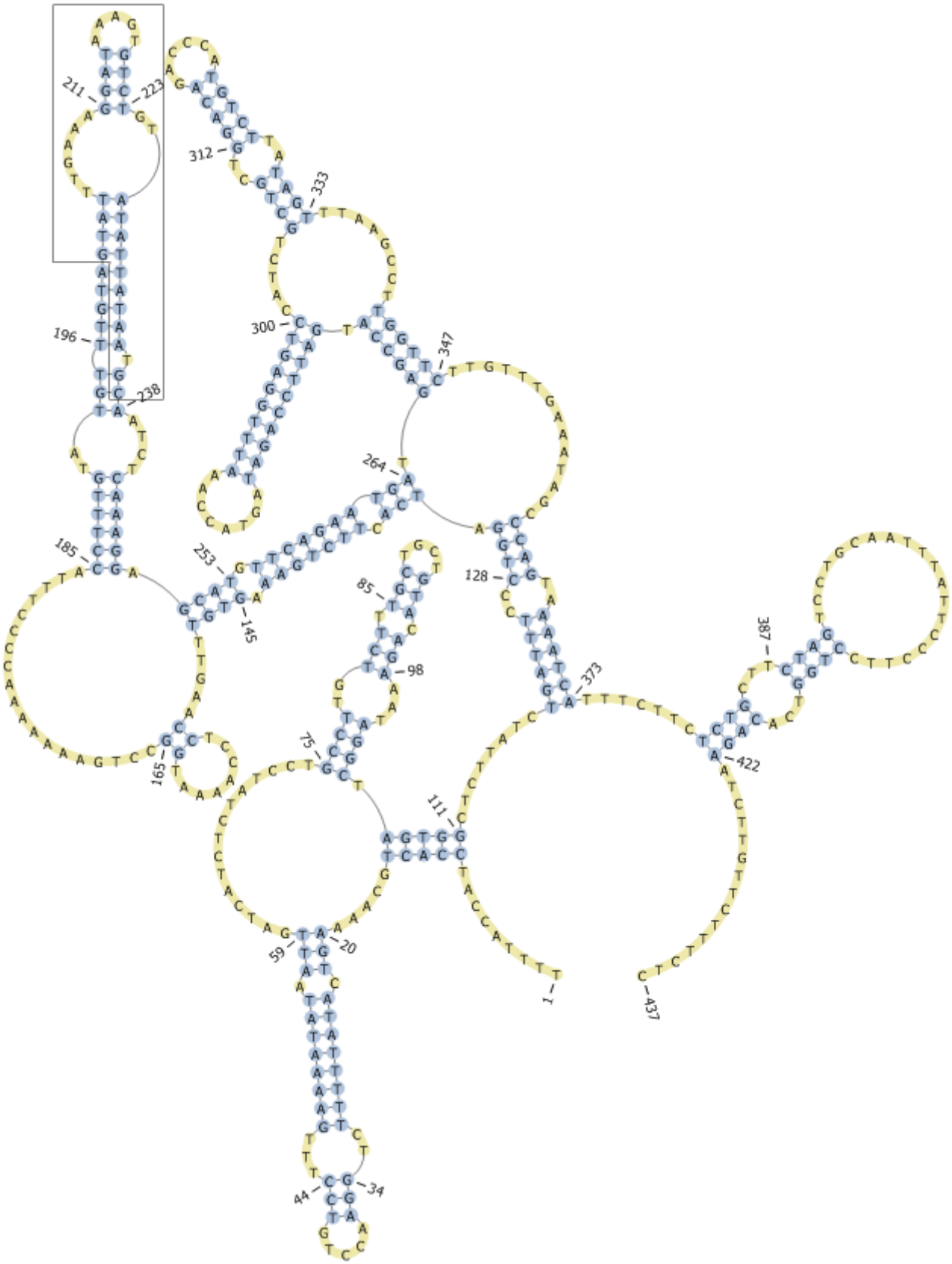


Abbildung 3.14: Unvollständiges Motiv

Ein Teil des Motivs aus *Pongo pygmaeus Assembly* PPYG2, Chromosom 1: 93565412 bis 93565448: -1 wird korrekt gefunden. Es fehlen jedoch weitere Bestandteile, wie das zweite Helix-*Internal Loop*-Helix Segment mit der darin enthaltenen Spaltstelle. Die eingerahmte Sequenz entspricht der *Pattern*-Beschreibung, faltet aber in eine andere Struktur. 3.14 wurde mit dem PseudoViewer Programm [Byun & Han, 2009] erzeugt und mit Gimp modifiziert.

3.2.2 Hepatitis Delta Virus Ribozyme

Da vorherige Suchen das Vorkommen Hepatitis Delta Virus (HDV) ähnlicher Ribozyme in Eukaryonten gezeigt haben [Ruminski et al., 2011], wurde am Beispiel der Amöbe *Dictyostelium discoideum* nach diesen Ribozymen gesucht. *Dictyostelium discoideum* ist ein Modellorganismus und für unsere Arbeitsgruppe von besonderem Interesse. Es wurden manuell nach verschiedenen RNAbob Deskriptoren (siehe Abschnitt 2.4.17) ohne anschließender Faltung gesucht. Die Suche ergab mit dem Suchmuster aus Abbildung 2.12(a) [Webb et al., 2009] keine Treffer. Auch das Entfernen von h1 (Abbildung 2.12(b)) oder die Verallgemeinerung von s4 (Abbildung 2.12(c)) ergaben keine Übereinstimmungen. Erst durch größere Verallgemeinerungen des Deskriptors konnten 19 bzw. 1761 Treffer identifiziert werden. Dazu wurde je ein *Mismatch* in r3, r4 und r5 (Abbildung 2.12(d)) zugelassen bzw. s3 mit 5 bis 8 N verallgemeinert (Abbildung 2.12(e)). Unter den Treffern befinden sich aufgrund der Sucheinstellung überlappende Sequenzen der gleichen Region. Es besteht die Gefahr, dass durch die Verallgemeinerung allein durch Zufall FP Sequenzen gefunden wurden. 110 der 1761 Treffer sind FP, da sie lange N Ketten besitzen und deshalb als passende Treffer zurückgegeben werden. 6 bzw. 413 der gefundenen Treffer werden mit Genen assoziiert und 1 Treffer mit dem Retrotransposon DIRS1 (Chromosom 2F:149224-149345:+). Tabelle 3.2 zeigt beispielhaft 4 der 6 nicht überlappenden HDV Ribozym ähnlichen Treffer.

[6]

Tabelle 3.2: HDV Ribozym Treffer

Chromosom	Strang	Start	Ende	ID	Name
2	+	2537842	2537988	DDB_G0273117	*ublcp1-1*
2	-	3493798	3493944	DDB_G0273861	*ublcp1-2*
2	-	2156416	2156578	DDB_G0272771	*wdr7*
5	-	4485438	4485597	DDB_G0290711	*DDB_G0290711_ps*

[6] Der Deskriptor in 2.12(d) wurde in vier *Dictyostelium discoideum* Genen gefunden.

3.2.3 Varkud Satellite Ribozym

Eine BLASTN Suche gegen die Nukleotiddatenbank von NCBI findet einen Teil der bekannten Sequenz in *Oryza latipes*, was jedoch mit einem E-Value von 2.5 nicht signifikant ist und Zufall sein kann. Es konnten keine weiteren Varkud Satellite ähnliche Ribozyme gefunden werden.

3.2.4 Hammerhead Ribozyme

Von allen gesuchten Ribozymen kommen *Hammerhead* Ribozyme am häufigsten vor, besonders als Typ I.

Da im konservierten, katalytischen Zentrum verschiedene Variationen beobachtet werden [Perreault et al., 2011], sind einige Deskriptoren an den Positionen 3 und 8 verallgemeinert.

Die Suche nach HHRz-Typ I, II und III zeigte vor allem in repetitiven Regionen aus *Xenopus tropicalis* einen modularen Aufbau, in dem alle drei Typen in einer Region möglich sind.

```
>Xenopus_tropicalis:JGI4.1:scaffold_8:[911491,911844]
TACCCTGGAACACAAGCCCCAAGAAGAACGAAACCGGTCTGTAGTTGGGAATCTGGTCT
                                (((((..(((((.............. Typ III Nr. 1
                                (((((..(((((.............. Typ III Nr. 2
ATTTGCGTACGTATGAGGCGGTCTCCTCAACCCTTTTGACTTGTTCGTGAATCCCCACT
............................................................ Typ III Nr. 1
............................................................ Typ III Nr. 2
CCTGACTCTACCTAAGCTGATCAGAACACCCTGCACTACACTGATGAGCCCCAAGAAGG
..............................))))))(....).((((......)) Typ III Nr. 1
............................................................ Typ III Nr. 2
                                                  (((..(( Typ II  Nr. 4
                           (((...(((((((....).((((......)) Typ I   Nr. 5
                                 ((.((((....).((((........ Typ I   Nr. 6
```

```
GCGAAACCGGTCTGTAGTTGGGAATCTTGACTTGTTTGTGAATCCCCACTCCTGGCTCT
))....))))                                                  Typ III Nr. 1
.....................................................       Typ III Nr. 2
      ((((..((((((.................................         Typ III Nr. 3
((....(..)..((((((.................................         Typ II  Nr. 4
))....(..)..))))))..)))                                     Typ I   Nr. 5
.....................................................       Typ I   Nr. 6
ACCTAAGCTGATCAGAACACCCTGCACTACACTGATGAGCCCCCAAGAAGGGCGAAACC
.......................))))))(....).((((.......))))....)) Typ III Nr. 2
.......................))))))(....).((((.......))))....)) Typ III Nr. 3
.......................))))))(....).)))).))))             Typ II  Nr. 4
.................................................))))....(. Typ I   Nr. 6
                          ((.((((....).((((.......))))....(. Typ I   Nr. 7
GGTCTGTTGTTAGGAATCTGATCAGCTATTATCCTGGCTTCTGCTTTAAAAACCCAGTG
))                                                          Typ III Nr. 2
))                                                          Typ III Nr. 3
.)..))).))                                                  Typ I   Nr. 6
.)..))).))                                                  Typ I   Nr. 7
```

Eine grafische Darstellung dieser Sequenz (Abbildung 3.15) verdeutlicht, dass je nach transkribiertem Abschnitt ein anderer Typ möglich ist. Auf dem Minusstrang konnte kein Ribozym gefunden werden.

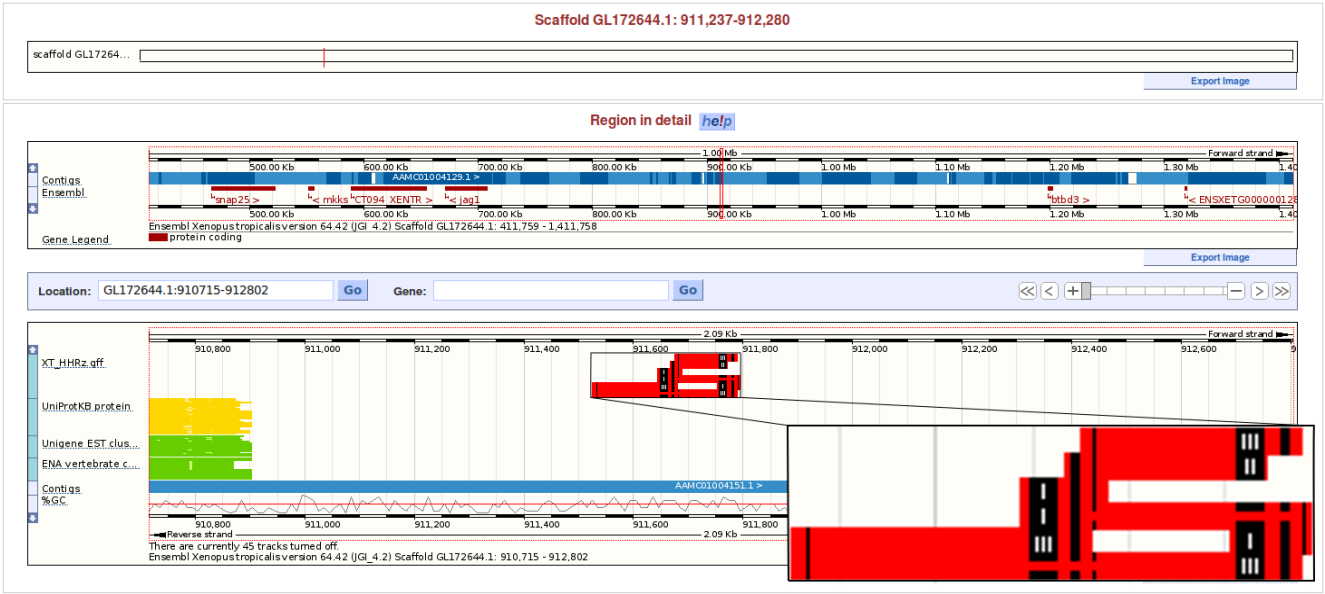


Abbildung 3.15: *Hammerhead* Ribozyme in einer genomischen Region von *Xenopus tropicalis*

Es gibt 7 mögliche HHRz aus *Xenopus tropicalis* im intergenischen Bereich auf dem Plusstrang von scaffold_8 Position 911491 bis 911844 (*Assembly* JGI4.1). Das Besondere an dieser Region ist, dass das konservierte, katalytische Zentrum des HHRz mehrmals vorkommt und somit alle drei HHRz-Typen möglich sind. Die roten Balken symbolisieren die transkribierte Sequenz. Schwarz markiert sind mit steigender Breite UH, GAAA und CUGANGA. Je nach Anordnung dieser „Bausteine“ ergeben sich die einzelnen Typen, wobei die Basenpaarung der angrenzenden Helices gewährleistet ist. An dieser Stelle befindet sich außerdem die Beschriftung der Typen. Die Markierung ist unterbrochen, um deutlich zu machen, welcher Typ gemeint ist.

Typ III

Tabelle 3.3 zeigt eine Übersicht der Suchergebnisse nach Typ III HHRz. Die Suchparameter sowie die verwendeten Deskriptoren befinden sich im Abschnitt 2.4.17.

7

Tabelle 3.3: *Hammerhead* Ribozym Typ III Ergebnisse

Datum	Suche	Faltung	Spezies	Primär	Filter
10. Jul. 2009	PatScan	UNAfold	191V 1A 631B 158E	0 (0V) 0 (0A) 426 (70B) 13063 (82E)	0 (0V) 0 (0A) 31 (20B) 718 (63E)
03. Sep. 2009	PatScan	keine	0V 0A 0B 3E	0V 0A 0B 174 (2E)	-
16. Sep. 2009	PatScan	UNAfold	0V 0A 0B 3E	0V 0A 0B 5644 (2E)	0V 0A 0B 444 (2E)
10. Nov. 2009	RNAbob	Mfold	269V 5A 806B 210E	420 (89V) 302 (4A) 18296 (508B) 561178 (160E)	15 (13V) 3 (2A) 923 (339B) 38725 (151E)
26. Mär. 2010	RNAbob	Mfold	1371V 4A 616B 174E	2336 (91V) 324 (4A) 19963 (509B) 600986 (162E)	430 (15V) 3 (2A) 924 (339B) 38755 (151E)
16. Jul. 2010	RNAbob	Mfold	0V 0A 0B *Arth* (1E)	0V 0A 0B 988 (1E)	0V 0A 0B 3 (1E)
05. Aug. 2010	PatScan	Mfold	*Viro* (1V) 0A 0B 0E	3348 (1V) 0A 0B 0E	87 (1V) 0A 0B 0E
16. Aug. 2010	PatScan	Mfold	0V 0A 0B 4E	0V 0A 0B 71 (3E)	0V 0A 0B 8 (3E)
27. Aug. 2010	RNAbob	Mfold	0V 0A 0B *Scma* (1E)	0V 0A 0B 1137 (1E)	0V 0A 0B 24 (1E)
02. Sep. 2010	PatScan	Mfold	3048V 50A 1222B 193E	476 (6V) 5 (5A) 85 (72B) 3259 (127E)	22 (2V) 0A 2 (2B) 42 (18E)
28. Sep. 2010	PatScan	UNAfold	0V 0A 0B *Cael* (1E)	0V 0A 0B 387 (1E)	0V 0A 0B 0E
04. Okt. 2010	PatScan	Mfold	0V 0A 0B *Xetr* (1E)	0V 0A 0B 105 (1E)	0V 0A 0B 67 (1E)
03. Nov. 2010	RNAbob	Mfold	0V 0A 0B 7E	0V 0A 0B 772 (6E)	0V 0A 0B 9 (3E)

Tabelle 3.3 – Fortsetzung

Datum	Suche	Faltung	Spezies	Primär	Filter
12. Nov. 2010	PatScan	Mfold	3048V 50A 1222B 150E	5 (4V) 11 (8A) 252 (173B) 7336 (124E)	0V 0A 2 (2B) 272 (39E)
03. Dez. 2010	PatScan	Mfold	3048V 50A 1222B 150E	2 (2V) 4 (4A) 193 (132B) 5897 (127E)	0V 0A 11 (3B) 31 (26E)
29. Apr. 2011	PatScan	Mfold	0V 0A 0B *Neve* (1E)	0V 0A 0B 2763 (1E)	0V 0A 0B 1 (1E)
18. Jul. 2011	PatScan	Mfold	0V 0A 0B *Xetr* (1E)	0V 0A 0B 13941 (1E)	0V 0A 0B 6 (1E)

[7] Übersicht der durchgeführten Suchen nach *Hammerhead* Ribozymen Typ III mit den jeweiligen Such- und Faltungsprogrammen, der Gesamtanzahl durchsuchter Spezies und der Anzahl gefundener sowie gefilterter Treffer. Das Balkendiagramm verdeutlicht in der „Spezies“ Spalte die Anzahl der Organismen pro Gruppe. Existiert nur ein Organismus, wie bei *Arabidopsis thaliana* (Arth), *Schistosoma mansoni* (Scma), *Caenorhabditis elegans* (Cael), *Xenopus tropicalis* (Xetr), *Nematostella vectensis* (Neve) oder eine Datei, wie bei den Satelliten RNA und Viroiden aus subviral DB (Viro), entspricht der Balken 100%. In den Spalten „Primär“ und „Filter“ steht das Diagramm für die Trefferanzahl in den Organismen pro Gruppe, wobei die Organismen in Viren (V), Archaeen (A), Bakterien (B) und Eukaryonten (E) unterteilt wurden. Viroide und Satelliten RNA wurden den Viren zugeordnet.

Die erste Anwendung der *Pipeline* in einer genomweite Suche nach HHRz III (10.07.2009) erfolgte mit PatScan und UNAfold. Aus mehr als 13000 primären Treffern wurden alle mehrfach vorkommenden und überlappenden Treffer ±5 nt gefiltert. Zusätzlich musste die Bedingung $\Delta G_{ratio} > 0.6 \frac{kcal}{mol}$ erfüllt sein. Somit konnten die ersten 749 möglichen HHRz III identifiziert und näher untersucht werden.

Am 03.09.2009 wurde anschließend mit den PatScan Deskriptoren 2.7(a) 2.8(b) und 2.9(b) nach allen drei Typen des HHRz in den Schistosomen gesucht, wobei die Positionen 3 und 8 des katalytischen Zentrums verallgemeinert und lediglich Watson-Crick Basenpaarungen erlaubt wurden. Durch das Zulassen von Wobble Basenpaaren wurden mit dem gleichen Deskriptor 2.9(b) 5470 weitere Treffer in den Schistosomen gefunden (16.09.2009). Durch die darauffolgende Faltung mit UNAfold und Filterung all jener Treffer, deren $\Delta G_{ratio} < 0.6 \frac{kcal}{mol}$ ist bzw. mit anderen Treffern ±5 nt überlappen oder mehrfach vorkommen, konnten 444 mögliche HHRz III bestimmt werden. Dieser Deskriptor wurde danach für eine zweite genomweite Suche (10.11.2009) verwendet. Bei dieser wurde jedoch mit einem anderen Such- (RNAbob) und Faltungsprogramm (Mfold) sowie mit zusätzlichen eigenen Faltungseinschränkungen gesucht, da Strukturen mit Helices aus einem oder zwei Basenpaaren von Mfold nicht berechnet werden können. Insgesamt wurden 39666 Treffer in allen Bereichen des Lebens vorhergesagt. Experimentell konnte Anne Kalweit, anhand der von Ihr als katalytisch aktiv getesteten Sequenzen einen weiteren Schwellenwert bestimmen, der in einer dritten genomweiten Suche (26.03.2010) als ein zusätzlicher Filter eingesetzt wurde. Treffer mit einer freien Energie $\Delta G_{free} > -10 \frac{kcal}{mol}$ wurden gefiltert, da diese wahrscheinlich nicht in die HHRz-Struktur falten. Außerdem wurde der $\Delta\Delta G$ Wert als Quotient durch die Differenz aus ΔG_{motif} und ΔG_{free} ersetzt [Seehafer et al., 2011]. Insgesamt konnten somit 40112 mögliche HHRz III identifiziert werden. Dieser Schwellenwert wurde weiter angepasst und bestätigte in *Arabidopsis thaliana* (16.07.2010) die zwei bisher bekannten HHRz III auf Chromosom 4 [Przybilski et al., 2005] und einen dritten neuen Treffer, der bereits von Anne Kalweit als katalytisch aktiv charakterisiert wurde. Außerdem wurden bei einer zweiten Kontrolle in den Sequenzen der subviral Datenbank [Rocheleau & Pelchat, 2006] 87 bekannte, einzigartige HHRz III bestätigt, bei denen die Motivbeschreibung übereinstimmte.

Bei der Suche vom 16.08.2010 wurde der Deskriptor vom 10.07.2009 in den *Loops* I und II um 100 nt vergrößert und mit ausschließlich festen Helixlängen in *Arabidopsis thaliana* und den Schistosomen gesucht. In *Arabidopsis thaliana* resultierte dies in den zwei bekannten Treffern. Die Treffer in den Schistosomen waren zu diesem Zeitpunkt unbekannt und wurden zum Teil von Perreault ein Jahr später bestätigt [Perreault et al., 2011].

Wird statt PatScan RNAbob verwendet (27.08.2010) und auf den minimalen ΔG_{free} Filter verzichtet, werden in *Schistosoma mansoni* 19 zusätzliche Treffer gefunden. Die Ausweitung des Suchraumes in einer vierten genomweiten Suche, inklusive aller Filterschritte und einer sehr strikten Wahl des $\Delta\Delta G$ Wertes ($\Delta\Delta G \leq 0 \frac{kcal}{mol}$), ergab in allen Genomen, bis auf den Archaeen, 66 mögliche HHRz III.

In einer wiederholten Suche in *Caenorhabditis elegans* (28.09.2010) mit dem Deskriptor, der bereits bekannte HHRz III fand, wurden 387 primäre Treffer gefunden, die nun mit UNAfold gefalten und den neuen Parametereinstellungen gefiltert wurden. Dabei wurden alle Treffer entfernt.

Die Verkleinerung von *Loop* II auf 4 bis 20 nt und Vergrößerung von *Loop* I auf bis zu 2000 nt ergab in *Xenopus tropicalis* 67 mögliche in *cis* oder *trans* spaltende HHRz III (04.10.2010). Um weitere in *trans* spaltende HHRz III zu ermitteln, wurden die *Loop*-Größen jeweils mit 100 bis 1000 nt definiert (12.11. und 03.12.2010) und in 4470 genomischen Sequenzen gesucht. *Loop* I Erweiterungen kommen bei gleichen Parametereinstellungen häufiger vor als *Loop* II Erweiterungen bei gleicher getesteter Wahrscheinlichkeit in Zufallssequenzen.

Anne Kalweit machte in ihren Experimenten eine weitere Beobachtung. Die häufig als Negativ-Kontrolle verwendete Mutante mit dem katalytischen Zentrum UG, CUGANGA, GAAA kann eine selbstspaltende Aktivität besitzen. Daraufhin wurde in 7 Eukaryonten nach dieser Mutation gesucht (03.11.2010). In 3 Eukaryonten konnten 9 einzigartige Sequenzen mit dieser Beschreibung identifiziert werden.

Die Suchen 29.04. und 18.07.2011 in *Nematostella vectensis* und *Xenopus tropicalis* zeigen weitere Anpassungen der Faltungs- und Filterparameter.

CUGANGA

Besonders hervorgehoben sei an dieser Stelle die genomweite Suche vom 26.03.2010, deren gefilterte C3G8 Sequenzen in der Publikation [Seehafer et al., 2011] veröffentlicht wurden. Die Suche in DNA-Sequenzen von 174 Eukaryonten, 616 Bakterien und 1371 Viren mit einer Gesamtgröße von rund $1,52 * 10^{11}$ Nukleotiden ergab 62836 Treffer mit einem C3G8 Basenpaar. Ungefähr 95% stammen aus Eukaryonten, 3% aus Viroiden und 2% aus Bakterien. Die Viroide enthalten bekannte HHRz [Tabler & Tsagris, 2004] und dienten als interne Kontrolle. Durch Anwendung der Analysepipeline (Abbildung 3.10) in der mit Hilfe thermodynamischer Parameter automatisch wahrscheinliche, katalytisch aktive HHRz bestimmt werden können, wurden zunächst alle Treffer mit einem ΔG_{free} Wert größer als -10 kcal/mol bestimmt (59932 Treffer) und anschließend alle mit einem $\Delta\Delta G$ Wert größer als 0.5 kcal/mol herausgefiltert, so dass letztendlich 858 Treffer verblieben. 284 der 858 Treffer sind einzigartige (*unique*) *Hammerhead* Ribozyme, von denen 122 aus 6 subviralen Pflanzenpathogenen und 2 aus *Arabidopsis thaliana* [Przybilski et al., 2005] bekannt sind. Die restlichen 156 sind mögliche neue Motive aus 50 Eukaryonten und 4 aus 3 Bakterien (*Azorhizobium caulinodans, Chloroflexus aggregans, Mycobacterium vanbaalenii*) (Tabelle 3.4).

8

Tabelle 3.4: *Hammerhead* Ribozyme Typ III

Gruppe	Spezies	Anzahl einzigartiger Motive
Primaten	*Homo sapiens*	2
	Pan troglodytes	1
	Macaca mulatta	3
	Microcebus murinus	3
	Tarsius syrichta	4
Nagetiere usw.	*Cavia porcellus*	2
	Mus musculus	3
	Ochotona princeps	2
	Oryctolagus cuniculus	2
	Rattus norvegicus	2
	Spermophilus tridecemlineatus	1
Laurasiatheria	*Bos Taurus*	2
	Canis familiaris	4
	Equus caballus	3
	Myotis lucifugus	1
	Sorex araneus	1

Tabelle 3.4 – Fortsetzung

Gruppe	Spezies	Anzahl einzigartiger Motive
	Sus scrofa	1
	Tursiops truncatus	1
	Vicugna pacos	2
Afrotheria	*Echinops telfairi*	4
	Loxodonta africana	1
Xenarthra	*Choloepus hoffmanni*	2
	Dasypus novemcinctus	2
andere Säugetiere	*Macropus eugenii*	1
	Monodelphis domestica	3
Vögel und Reptilien	*Gallus gallus*	1
	Taeniopygia guttata	2
Amphibien	*Xenopus tropicalis*	8
Fische	*Danio rerio*	2
	Tetraodon nigroviridis	1
Insekten	*Aedes aegypti*	1
	Culex quinquefasciatus	2
	Drosophila persimilis	3
	Drosophila pseudoobscura	5
	Nasonia vitripennis	2
	Tribolium castaneum	2
Pflanzen	*Arabidopsis lyrata*	3
	Arabidopsis thaliana	3
	Physcomitrella patens	1
	Vitis vinifera	2
Pilze	*Aspergillus flavus*	1
	Aspergillus oryzae	1
	Pichia stipitis	1
andere Metazoa	*Caenorhabditis remanei*	1
	Caenorhabditis briggsae	1
	Hydra magnipapillata	35
	Ixodes scapularis	1
	Schistosoma japonicum	2
	Schistosoma mansoni	24
Bakterien	*Azorhizobium caulinodans*	2
	Chloroflexus aggregans	1
	Mycobacterium vanbaalenii	1

[8] Übersicht der 160 neuen HHRz Typ III Sequenzen aus [Seehafer et al., 2011], gruppiert nach der Ensembl Klassifizierung und visualisiert durch das Balkendiagramm.

Die meisten Treffer befinden sich in *Hydra magnipapillata* (35) (siehe Tabelle 3.5) und *Schistosoma mansoni* (24).

[9]

Tabelle 3.5: *Hammerhead* Ribozyme Typ III aus *Hydra magnipapillata* und *Xenopus tropicalis*

Spezies	Motivname	Sequenz	Region	verfügbare Expressions-daten	Häufigkeit im Genom
Hydra magni-papillata	Hyma2	TGATTGTCCATGTCCGGAATAAAT ATGTTCCTGGACATGCTGATGAGC CCTGATATTGGGCGAAACAATCA	intergenisch	keine	32

Tabelle 3.5 – Fortsetzung

Spezies	Motivname	Sequenz	Region	verfügbare Expressions-daten	Häufigkeit im Genom
	Hyma8	TGATTGTCCATGTCCAGAATAAAT ATGTTCCTAGACATGCTGATGAGC CCTGATATTGGGCGAAACAATCA	Repeat hAT-2_NV, hAT-10_NV, none13881, none2351	keine	1
	Hyma14	GATTGTCCATGTCTGGAATGAAAT ATGTTCCTAGACATGCTGATGAGC CCTGATATTGGGCGAAACAATC	Repeat hAT-2_NV	keine	1
	Hyma19	ATTGTCCATGTCCGGAATAAATAT GTTCCTAGACATGCTGATGAGCCC TGATATTGGGCGAAACAAT	Repeat hAT-2_NV, hAT-10_NV, none13881, none2351	keine	1
	Hyma20	TGATTGTCCATGTCCGGAATAAAT AGTTCCTAGACATGCTGATGAGCC CTGATATTGGGCGAAACAATCA	intergenisch	keine	1
	Hyma22	TGATTGTCCATGTCTGCAATGAAA TATGTTCCTAGACATGCTGATGAG CCCTGATATTGGGTGAAACAATCA	intergenisch	keine	1
	Hyma25	GGTTGTCCGTGTCTGGAATGAAAT ATGTTCCTAGACATGCTGATGAGC CCTGATATTGGGCGAAACAATC	Repeat hAT-2_NV	keine	3
	Hyma31	TGATTGTCCATGTCCGGAATAAAT ATGTTCTTAGACATGCTGATGAGC CCTGATATTGGGCGAAACAATCA	Repeat hAT-2_NV, hAT-10_NV, none13881, none2351	keine	1
	Hyma32	TGATTGTCCATGTCCGGAATAAAT ATGTTCCTAAACATGCTGATGAGC CCTGATATTGGGCGAAACAATCA	intergenisch	keine	2
	Hyma34	GATTGTCCATGTCCGGAATAAATA TGTTCCTAGACATGCTGATGAGCT CTGATATTGGGCGAAACAATC	Repeat hAT-2_NV, hAT-10_NV, none13881, none2351	keine	1
Xenopus tropicalis	Xetr1	ACCGGTCTGTAGTTGGGAGTCTGA TCAGTTATTATCCTGGCTGTTGCT TCAAAACTACACTGATGAGCCCCA AGAAGGGCGAAACCGGT	intronisch	CX838854, DN033733	1
	Xetr2	CCGGTCTGTAGTTGGGAATCTGAT CAGCTATTATCCTGGCTTTTGCTT CAAAAAGCTAAGCTGATCAGAAAA CCCTGCACTACACTGATGAGCCCC AAGAAGGGCGAAACCGG	intronisch	CX741003, CR576111	1
	Xetr3	TGTGTCTGCAGCTGCCCACGTGCC AGAGAGGGTTGTGCCCATTCCCCT CGTGGGCCGCGCTGCACTGATGAG TCCCAAACAGGACGAAACACA	intergenisch	keine	1
	Xetr4	GCCGGTCTGTAGTTGGGAATTTGT ATGATGCAGTTTCCTCAACCCTAA TGATTTGTTTGTGAATTCACACTT ATGGCTCTACCTAAGCTGATCAGA AAACACTGCACTACACTGACGAGC CCCAAAAAGGGCGAAACTGGT	intronisch	ENSXETT 00000020052	3

Tabelle 3.5 – Fortsetzung

Spezies	Motivname	Sequenz	Region	verfügbare Expressions-daten	Häufigkeit im Genom
	Xetr5	GTGTCTGCAGCTGCCCACGTGCCA GAGAGGGTTGTGCCCATTCCCCTC GTGGGCCGCGCTGCACTGATGAGT CCCAAACAGGACGAAACAC	intronisch	ENSXETG 00000010228	1
	Xetr6	AGTTTGCCTTCACATGTACTTTTA GTATAAAACATGAATGTAAAAAGG CCTGAAGAGAAAGATAAGTAATAA AAAGTAACAATAACGAAAAATCTG AACTGCAAAGATTAATAGTTTTTT GGCTGCTGGAGGAGTTTTTTGAAA ACT	intronisch	X920536	1
	Xetr7	ACCGGTCTGTAGTTGGGAATCTTG ACTTGTTTGTGAATCCCCACTCCT GGCTCTACCTAAGCTGATCAGAAC ACCCTGCACTACACTGATGAGCCC CAATAAGGGCGAAACCGGT	intronisch	ENSXETG 00000025944	2
	Xetr8	CCGGTCTGTAGTTGGGAATCTTGA CTTGTTTGTGAATCCCCACTCCTG GCTCTACCTAAGCTGATCAGAACA CCCTGCACTACACTGATGAGCCCC CAAGAAGGGCGAAACCGG	intergenisch	keine	1

[9] 3.5 zeigt 10 der 35 möglichen HHRz aus *Hydra magnipapillata* (*Assembly* h7) und 8 aus *Xenopus tropicalis* (*Assembly* JGI4.1). Die gesamte Tabelle befindet sich im *Supplemental Material* der Publikation [Seehafer et al., 2011].

Sara Völkel und Anne Kalweit konnten experimentell nachweisen, dass Hyma2, Hyma14, Hyma19, Hyma20, Hyma22, Hyma25, Hyma31 (Völkel) und Xetr1, Xetr2, Xetr4, Xetr5, Xetr8 (Kalweit) aus Tabelle 3.5 katalytisch aktiv sind.

Die Annotation erfolgte mit Hilfe der Skripte getEST und getSupportingEvidence (Abschnitt 3.1.1). Einige der 160 Treffer wurden zeitgleich in einer Veröffentlichung von De la Peña *et al.* gefunden [de la Peña & Garcia-Robles, 2010b]. Bis auf wenige Ausnahmen befinden sich die Treffer in repetitiven Sequenzbereichen, in Introns proteinkodierender Gene oder in intergenischen Regionen. Auffallend ist, dass innerhalb der Organismen zum Teil nur ein HHRz III Motiv gefunden wurde, welches jedoch mehrmals vorkommen kann [Seehafer et al., 2011]. Das Vorkommen in repetitiver Satelliten DNA von Egel [Ferbeyre et al., 1998], Grillen [Rojas et al., 2000] und Amphibien [Epstein & Gall, 1987] sowie in Introns Proteinkodierender Gene des Menschen [de la Peña & Garcia-Robles, 2010a] wurde bereits zuvor beschrieben. Einige der HHRz aus intergenischen Regionen stehen im Zusammenhang mit transposablen Elementen, was ebenfalls von De la Peña *et al.* beobachtet wurde [de la Peña & Garcia-Robles, 2010b].

Die folgenden Treffer in den *Arabidopsis* wurden durch manuelle Homologiesuchen mit BLAST ermittelt. Das HHRz „Ara1“ in *Arabidopsis thaliana* auf dem Minusstrang von Chromosom 4 an der Position 15027834 bis 15027890 [Przybilski et al., 2005] besitzt ein Homolog in *Arabidopsis lyrata* auf dem Minusstrang von Chromosom 7 an der Position 4574246 bis 4574297. Anne Kalweit konnte zeigen, dass das Transkript dieser Sequenzregion („Arly3“) sich ebenfalls selbst spalten kann. Ein Homolog zu „Ara2“ existiert in *Arabidopsis lyrata* nicht. Die Sequenz wird der gleichen Position wie „Ara1“ zugeordnet (siehe Abbildung 3.16). „Ara1“ liegt im intergenischen Bereich zwischen den Genen *ASHR3* und *ATMUS81*, die für eine Histone-lysine N-methyltransferase und eine Endonuklease kodieren. Die Proteine und deren Funktionen im Umfeld der homologen Position in *Arabidopsis lyrata* sind unbekannt. *XM_002869310.1* kodiert für ein hypothetisches Protein. Abbildung 3.17 zeigt die *Synteny* zwischen *Arabidopsis thaliana* und *Arabidopsis lyrata*. Teile des Chromosoms 4 aus *Arabidopsis thaliana* können in *Arabidopsis lyrata* den Chromosomen 6 und 7 zugeordnet werden.

Das neu gefundene HHRz III Arth1 (26.03.2010) von Chromosom 1:23584856-23584979:+ besitzt eine homologe Sequenz in *Arabidopsis lyrata* auf Chromosom 2:798764-798888:+ *Assembly* Araly1.1 (Abbildung 3.18). Falls diese katalytisch aktiv ist, handelt es sich um eine weitere Variation des HHRz.

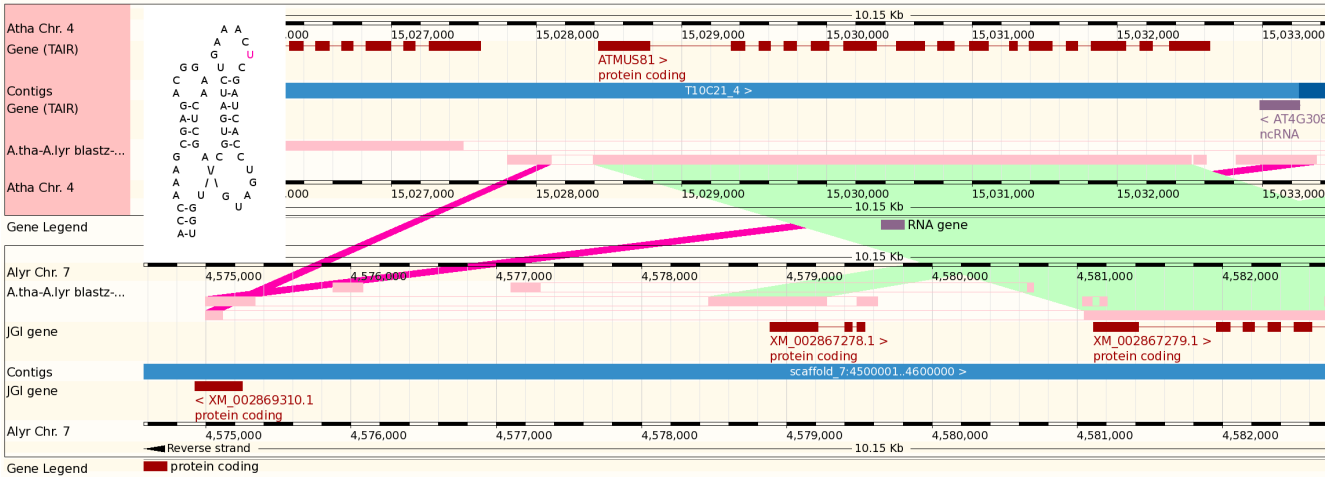


Abbildung 3.16: *Arabidopsis Hammerhead* Ribozyme

3.16 zeigt einen Bildausschnitt der Ensemblgenomes Webseite. Zu sehen ist ein „*Multi-Species View*“, in dem die Sequenzregion von Chromosom 4 aus *Arabidopsis thaliana* (oben) mit der Sequenzregion von Chromosom 7 aus *Arabidopsis lyrata* (unten) verglichen wird. Die manuell in mangenta eingezeichneten Balken zeigen das *Mapping* der HHRz-Regionen auf die gleiche Position in *Arabidopsis lyrata*. Dies wird durch die Suche vom 02.09.2010 bestätigt. Zusätzlich ist die Sekundärstruktur des in *Arabidopsis thaliana* gefundenen HHRz „Ara1“ im Vergleich zu dem neu gefundenen HHRz in *Arabidopsis lyrata* „Arly3“ dargestellt. Der Unterschied ist das in mangenta gefärbte Uridin, welches in *Arabidopsis lyrata* nicht vorhanden ist.

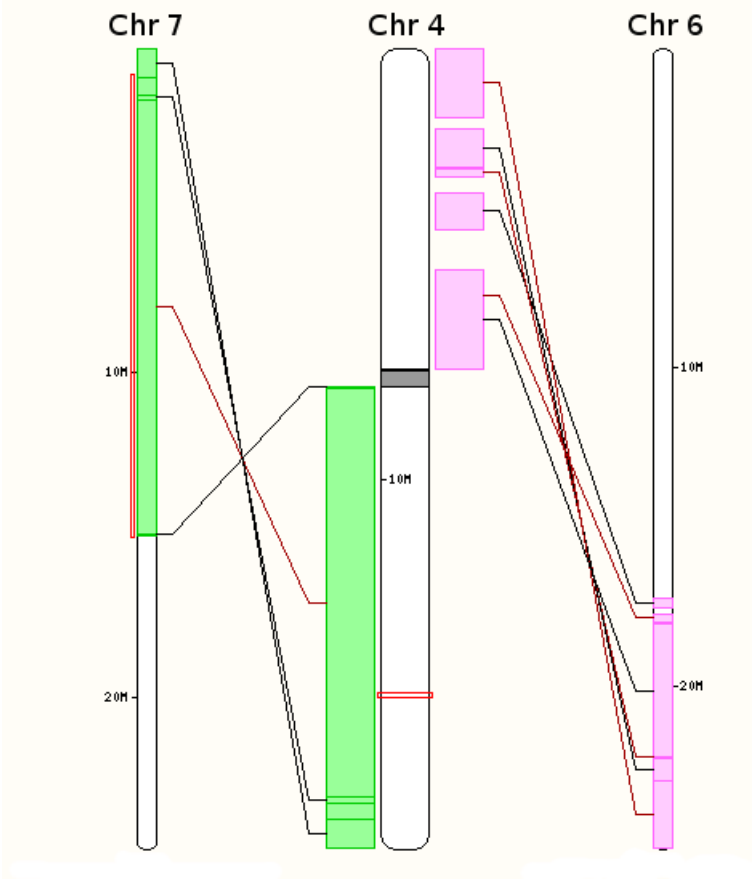


Abbildung 3.17: *Arabidopsis Synteny*

3.17 zeigt das Ergebnis eines globalen *Alignments* des gesamten *Arabidopsis thaliana* Genoms (AT) gegen das *Arabidopsis lyrata* Genom (AL). Chromosom 4 (AT) wurde auf Chromosom 6 und 7 (AL) abgebildet. 3.17 wurde von Ensemblgenomes erstellt, übernommen und modifiziert.

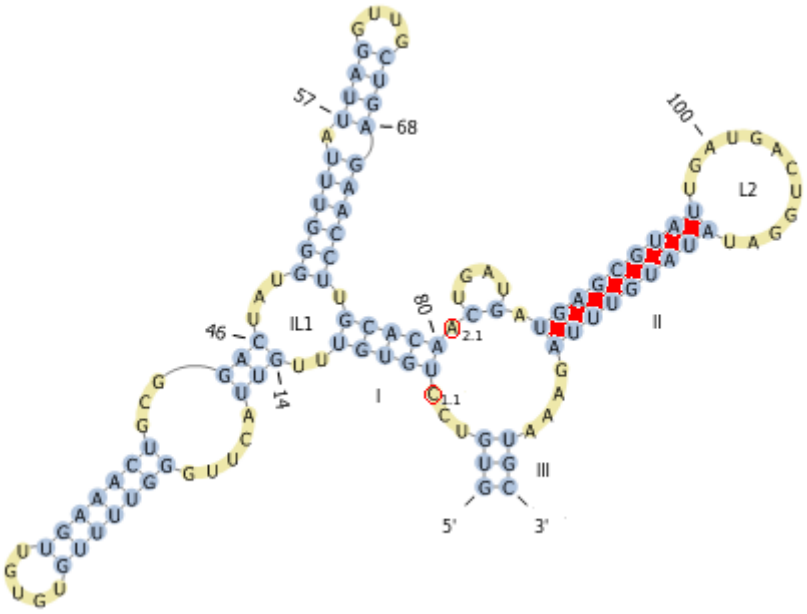


Abbildung 3.18: *Arabidopsis Hammerhead* Ribozym Variation
Das dargestellte *Hammerhead* Ribozym entspricht, falls es katalytisch aktiv ist, einer weiteren Variation des Motivs. Diese besteht aus einem C1.1A2.1 *Mismatch* (ähnlich wie in *Schistosoma haematobium* [Ferbeyre et al., 1998]) und einem 9 Basenpaar langen Stem II. 3.18 wurde mit dem PseudoViewer erstellt und mit KolourPaint modifiziert.

Das Erstellen eines MSA ist aufgrund der variablen Sequenzlängen der gefundenen Treffer schwierig. Außerdem wird, wie unter Abschnitt 3.1.1 erwähnt, das konservierte, katalytische Zentrum, welches in jeder Sequenz enthalten ist, nicht erkannt. Um dennoch einen phylogenetischen Baum erstellen zu können, wurden zwei auf Distanzen basierende Ansätze verfolgt, die zu ähnlichen Bäumen führten (Abbildung 3.20 und 3.21). Zum einen wurde mit Hilfe des Skriptes getL1L2.pl und ClustalW ein MSA erzeugt, in dem die *Loop*-Regionen bis auf 4 flankierende Nukleotide gekürzt, die Teilsequenzen blockweise aligniert und zu einem *Alignment* zusammengefügt wurden. Anschließend wurde das *Alignment* manuell korrigiert, um eine größere Sequenzähnlichkeit zu erreichen. Abbildung 3.19 zeigt einen Ausschnitt dieses *Alignments*.

Danach wurde mit dem Neighbor Joining Algorithmus und 1000 Bootstrap-Wiederholungen ein ungewurzelter Baum generiert (Abbildung 3.20). Die Sequenzen eines Organismus konnten deutlich in Cluster unterteilt werden (z. B. *Hydra magnipapillata*, *Schistosoma mansoni* oder *Tarsius syrichta*). Auch Cluster bestehend aus *Arabidopsis thaliana* und *Arabidopsis lyrata*, *Drosophila persimilis* und *Drosophila pseudoobscura* oder *Aspergillus flavus* und *Aspergillus oryzae* wurden in Gruppen zusammengefasst und bilden Gattungen. Es existieren ebenso Gruppierungen von Klassen, wie z. B. *Gallus gallus* und *Taeniopygia guttata*, die eine Konservierung innerhalb der Vögel darstellen. Dies bestätigt eine Homologiesuche im neu verfügbaren Genom *Meleagris gallopavo*, wo sich ein mögliches HHRz im intergenischen Bereich auf Chromosom 6:32863677-32863825:- (*Assembly* Turkey_2.01) befindet. Des Weiteren besteht, den Bootstrap-Werten zu Folge, zwischen den Treffern des Reismehlkäfers (*Tribolium castaneum*) und des Krallenfrosches (*Xenopus tropicalis*) eine höhere Konservierung als zwischen den Pflanzen *Arabidopsis thaliana* und *Vitis vinifera*.

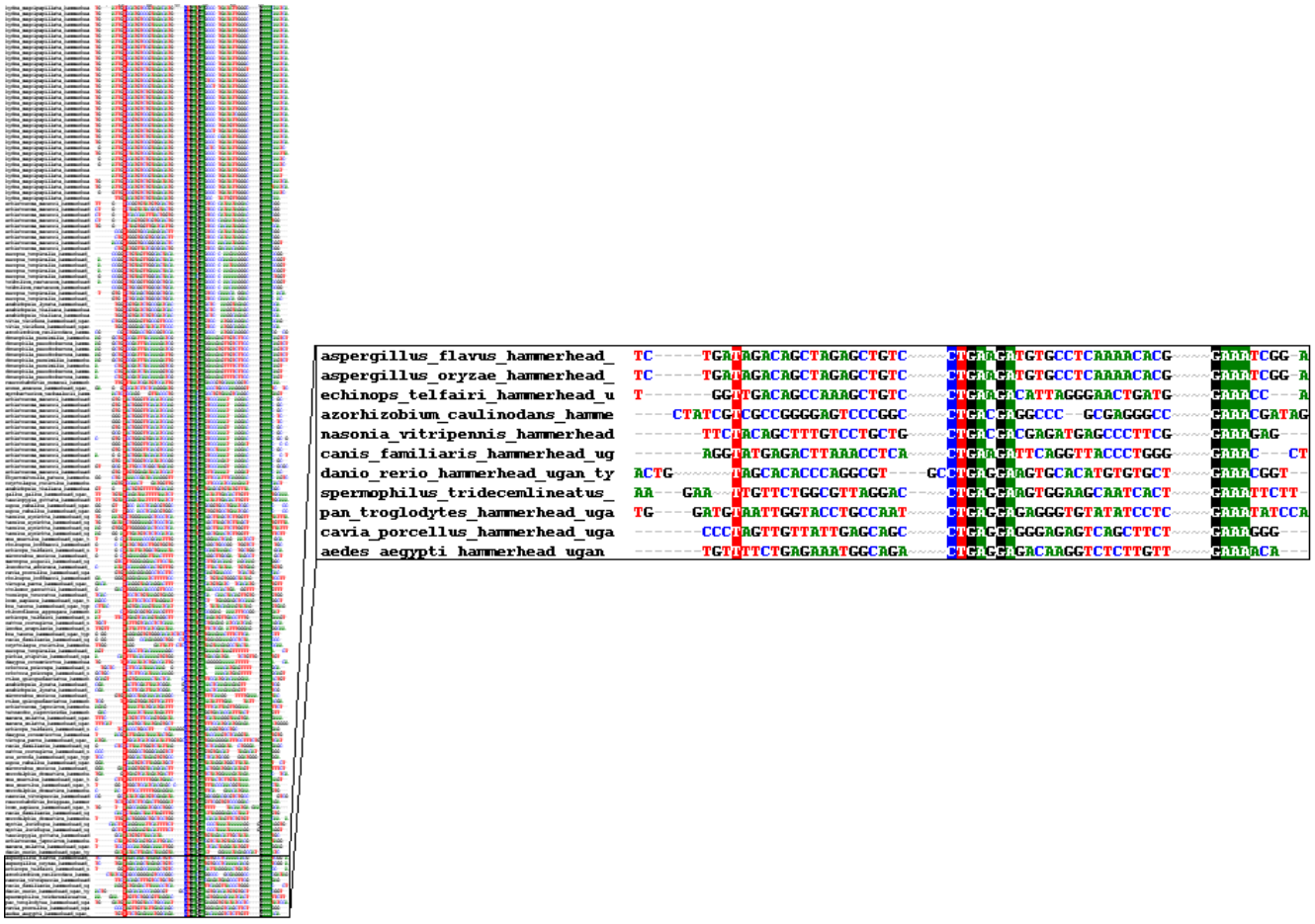


Abbildung 3.19: *Multiple Sequence Alignment*
Das erstellte MSA der 160 gefundenen, neuen möglichen *Hammerhead* Ribozyme Typ III zeigt in rot und dunkelgrün die Spaltstelle und das konservierte, katalytische Zentrum.

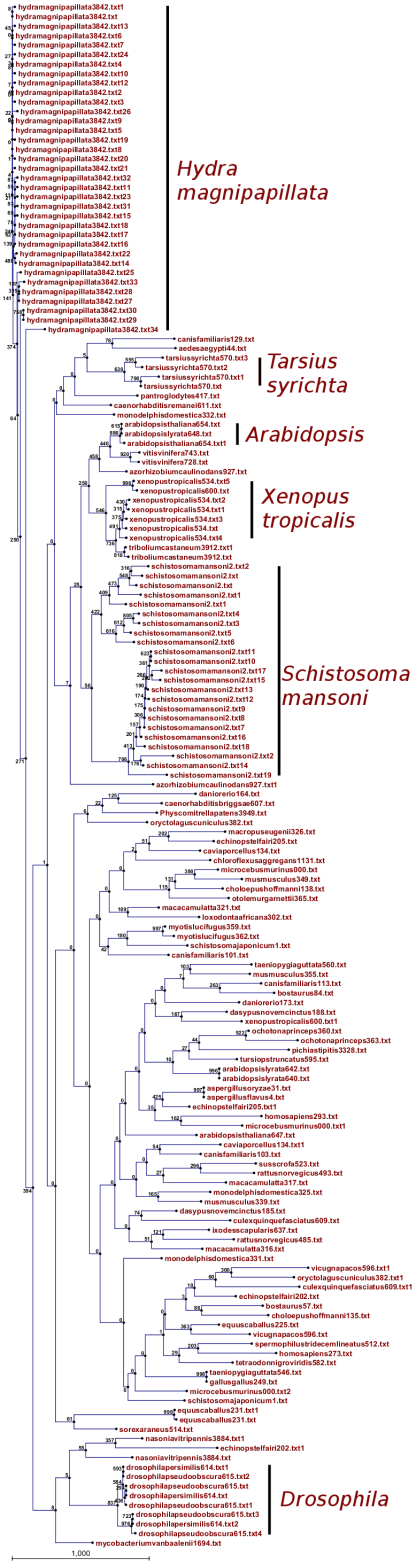


Abbildung 3.20: Phylogenetischer Baum

Der phylogenetische Baum der 160 HHRz Typ III Kandidaten wurde mit dem Neighbor Joining Algorithmus erzeugt. Die Kantenlängen entsprechen, ähnlich wie in der *Heatmap* (Abbildung 3.21), Distanzen zwischen den Sequenzen. Die Zahlen an den Knoten verdeutlichen, wieviel der Bootsrap Bäume ebenfalls diese Kanten besaßen. Es sind Gruppierungen innerhalb der Sequenzen eines Organismus erkennbar, aber auch Cluster von Gattungen und Klassen.

Der andere Ansatz ist das Berechnen der Levenshtein-Distanz zwischen den Sequenzen und die Erstellung einer Abstandsmatrix. Diese kann z. B. mit Hilfe einer *Heatmap* visualisiert werden. Nach der Sortierung der Werte ergibt sich der in Abbildung 3.21 dargestellte Baum [Seehafer et al., 2012].

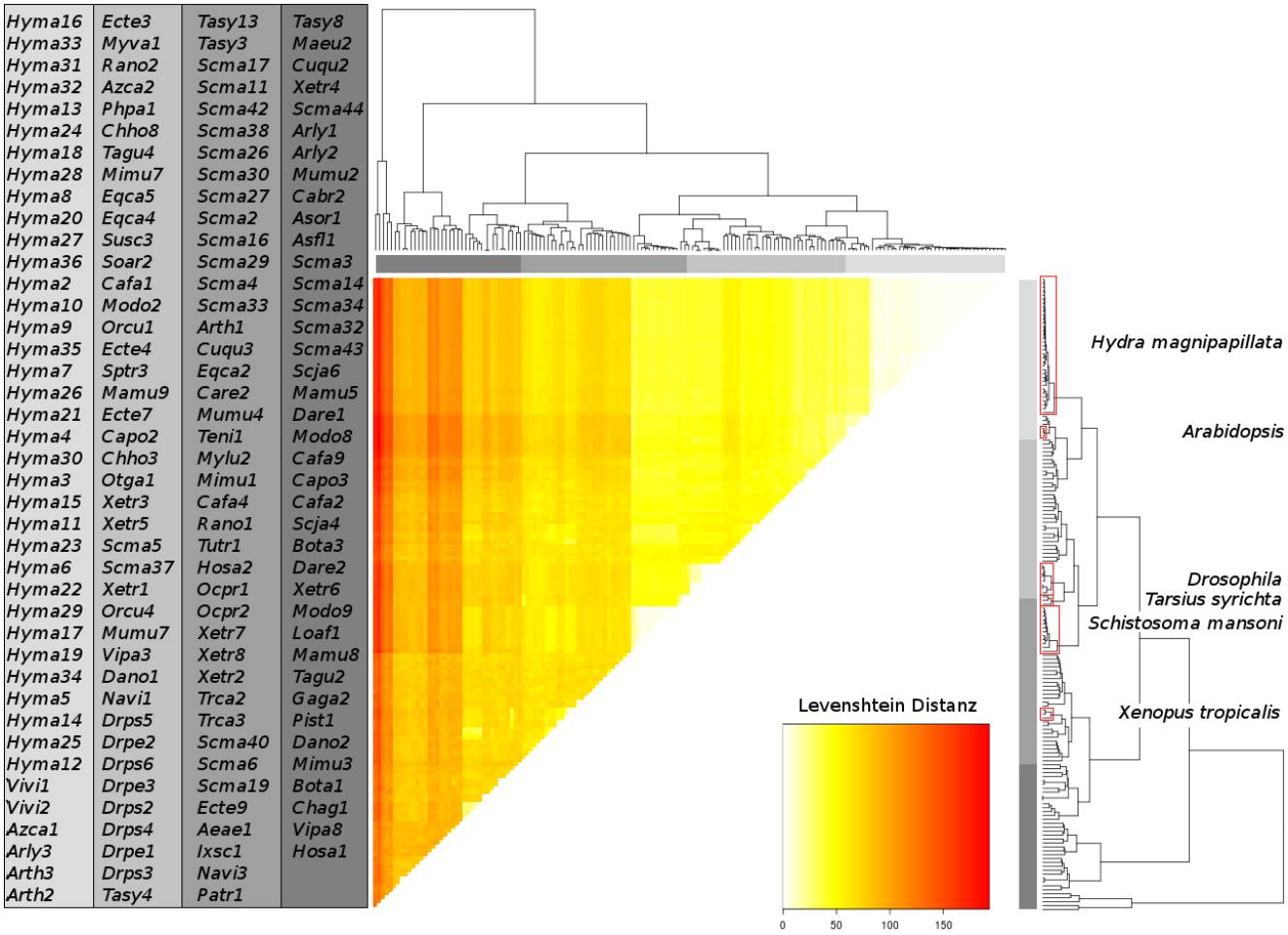


Abbildung 3.21: Sortierte *Heatmap*

Diese *Heatmap* aus [Seehafer et al., 2012] visualisiert die sortierte, symmetrische Levenshtein-Distanzmatrix der 160 Treffersequenzen. Rot entspricht dabei einem maximalen und weiß einem minimalen Abstand. Der minimale Abstand der *Hydra magnipapillata* Sequenzen untereinander ist gut erkennbar. Die größte Distanz mit 193 *Edit*-Operationen besteht zwischen *Mycobacterium vanbaalenii* (Myva1) und *Vicugna pacos* (Vipa8). Die rot umrandeten Teilbäume zeigen beispielhaft die Gruppierung in Organismen und Gattungen.

Die Gruppierung in Organismen z. B. bei *Hydra magnipapillata* stimmt mit dem Baum in Abbildung 3.20 überein. Ebenso die Treffer aus *Drosophila persimilis* und *Drosophila pseudoobscura* wurden in einem Teilbaum als Gattung *Drosophila* zusammengefasst. Des Weiteren wurde die Homologie zwischen den Treffern des Reismehlkäfers (*Tribolium castaneum*) und des Krallenfrosches (*Xenopus tropicalis*) bestätigt. Eine weitere Übereinstimmung ist ein HHRz im Bakterium *Azorhizobium caulinodans*, das eine große Ähnlichkeit zu den Treffern in *Arabidopsis thaliana*, *Arabidopsis lyrata* und *Vitis vinifera* aufweist und sich deshalb im gleichen Cluster befindet. Ein Unterschied zwischen den Bäumen ist die Verteilung der *Schistosoma mansoni* Treffer auf 5 Cluster.

Die Suche nach *Hairpin* Ribozymen 200 nt vor oder hinter den 160 HHRz III, wie sie bei einigen Viroiden und Satelliten RNA gefunden werden, ergab auf beiden Strängen keine Treffer.

Unter Berücksichtigung der in der Einleitung Abschnitt 1.6 definierten Formeln berechnet sich die Wahrscheinlichkeit des Deskriptors aus Abbildung 2.9(a) ohne die alternativen Positionen wie folgt:

$P_{HelixIII} = \left(\frac{6}{16}\right)^3$

$P_{HelixI} = \left(\frac{6}{16}\right)^4$

$P_{HelixII} = \left(\frac{6}{16}\right)^4$

$P_{LoopI} = 1^4$

$P_{LoopII} = 1^4$

$P_{GAAA} = \left(\frac{1}{4}\right)^4$

$P_{UH} = \frac{1}{4} * \frac{3}{4}$

$P_{CUGANGA} = \left(\frac{1}{4}\right)^6 * 1$

$P = P_{HelixIII} * P_{HelixI} * P_{HelixII} * P_{LoopI} * P_{LoopII} * P_{GAAA} * P_{UH} * P_{CUGANGA} = 3.687613 * 10^{-12}$

Das heißt in einer Zufallssequenz der Länge $2.72 * 10^{11}$ wird das Motiv 1 mal erwartet.

Durch die variablen *Loop*- und Helixnukleotide ist jedoch die Wahrscheinlichkeit einen Treffer zu finden höher. Aus diesem Grund wurden mit Hilfe des Skriptes getObservedMotifProbability 5 Zufallssequenzen der Länge $2.2 * 10^8$ erzeugt und darin mit dem Deskriptor gesucht. Durchschnittlich wurden 8 Treffer pro Strang ermittelt, was einer Motivwahrscheinlichkeit von $3.63 * 10^{-8}$ entspricht. In den durchsuchten Sequenzen werden also 10 mal mehr Treffer gefunden als die 5518 erwarteten Treffer in Zufallssequenzen.

Zur Abschätzung der optimalen Parametereinstellungen wurde eine ROC-Kurve erstellt (Abbildung 3.22). 50% der getesteten Kombinationen ergaben eine leere Treffermenge aufgrund zu strikter Filterbedingungen.

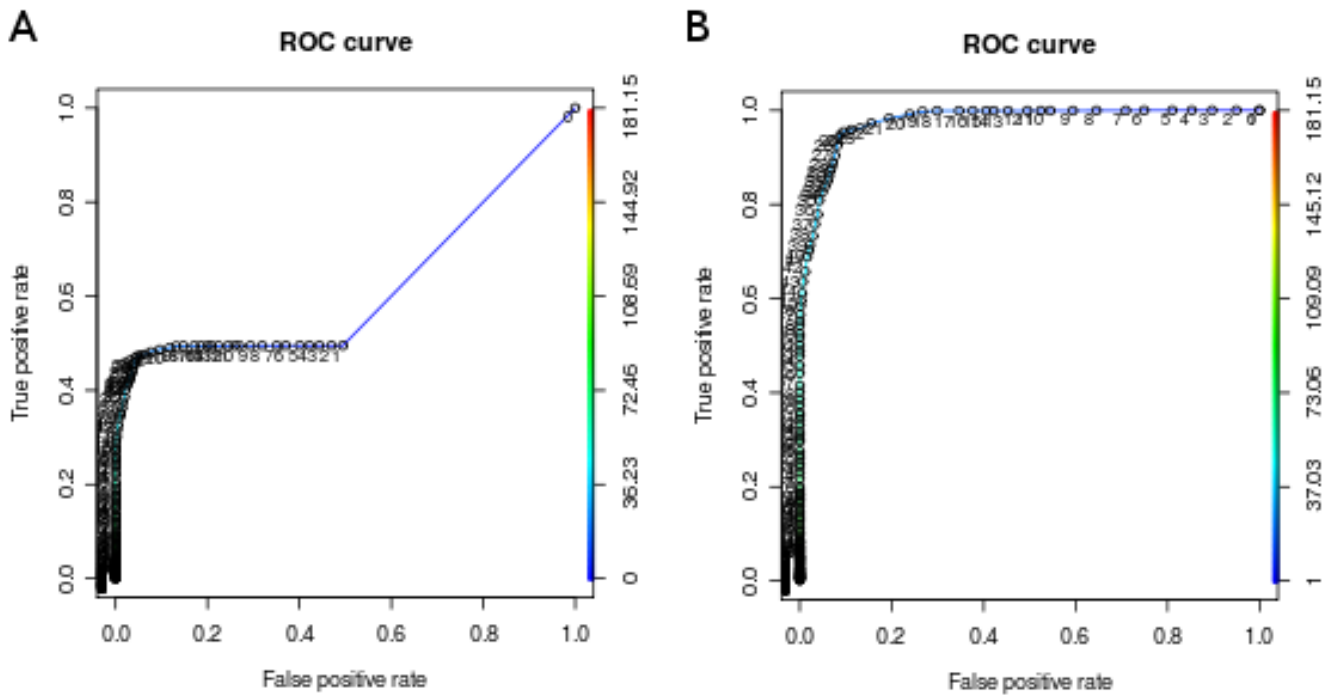


Abbildung 3.22: ROC-Kurve
Die ROC-Kurve zeigt jeweils die Rate der True Positives (Sensitivität) = True Positive / (True Positive + False Negative) aufgetragen gegen die Rate der False Positives = False Positive / (False Positive + True Negative). In B) sind im Gegensatz zu A) alle Scores herausgefiltert, bei denen weder TP noch FP gefunden wurden, was in 50% aller getesteten Parameterkombinationen der Fall war.

Am Verlauf der ROC-Kurve ist zu sehen, dass eine der besten Parametereinstellungen bei 26 TP und 0 FP erreicht wird. Aus der Ergebnis-Datei des Programms ROCscript.pl kann anhand der Zeilennummer die Parametereinstellung ermittelt werden. RNAhit wurde mit einem ΔG_{free} Wert von -19 kcal/mol, einem $\Delta\Delta G$ Wert von 2 kcal/mol, einem Überlapp-Wert von 0 sowie unter Verwendung des Unique-Filters, UNAfold und RNAbob aufgerufen.

Ein Teil der Treffer wurde von Anne Kalweit experimentell validiert, insbesondere die Sequenzen aus dem afrikanischen Krallenfrosch *Xenopus tropicalis*. Dazu erstellte Sie mit rekursiver *Polymerase Chain Reaction* (PCR) aus überlappenden DNA-Oligonukleotiden künstliche DNA-Matrizen, von denen durch *in vitro* Transkription RNA erhalten werden konnte und die anschließend auf katalytische Aktivität getestet wurden. Das Beispiel in Abbildung 3.23 ist Xetr8 aus dem *Assembly* JGI4.1 scaffold_8: 911673 - 911788 :1. Die Sequenz entspricht Nummer 3 aus dem Beispiel des modularen Aufbaus Seite 65.

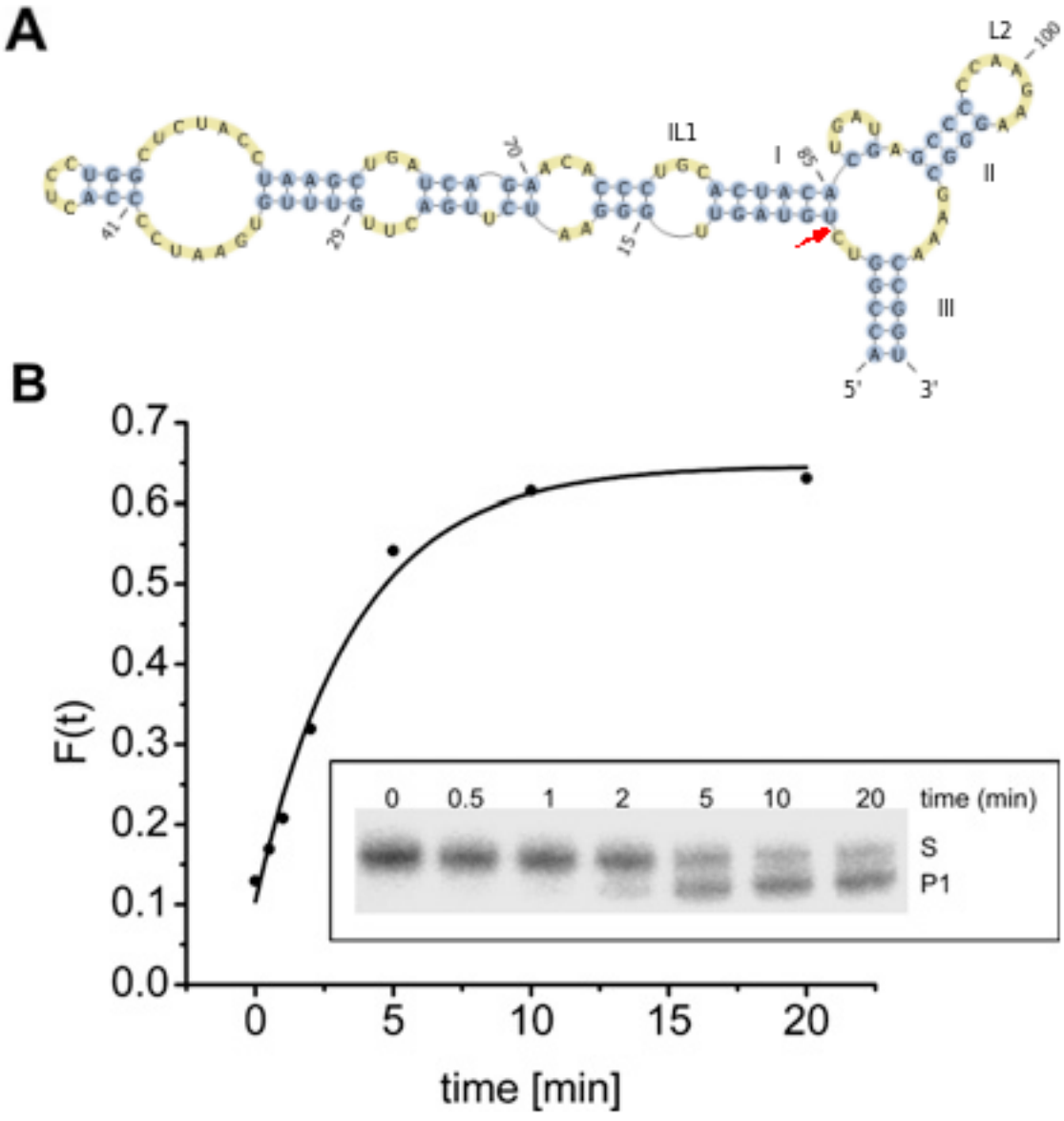


Abbildung 3.23: *Hammerhead* Ribozym Typ III Xetr8 aus *Xenopus tropicalis*
A) zeigt die Sekundärstruktur der Sequenz mit der rot markierten Spaltstelle sowie die interagierenden *Loops* IL1 und L2. B) zeigt die Verlaufskurve der Selbstspaltungsreaktion. Das Gelbild veranschaulicht das Verschwinden des Substrates S und das Erscheinen des 3' Produktes P1. Die ermittelte Spaltungsrate k_{obs} [Stage-Zimmermann & Uhlenbeck, 1998] beträgt 0.28 min^{-1}. 3.23 wurde modifiziert aus [Seehafer et al., 2011] übernommen.

In Zusammenarbeit mit Franziska Hoffgaard konnten mit Hilfe der *Mutual Information* (MI) zwischen den Helices und innerhalb der Helices, je nach Typ, signifikante co-evolutionäre Zusammenhänge gezeigt werden. Diese Zusammenhänge bestehen besonders bei Nukleotidpositionen zwischen den Helices I und III sowie innerhalb der Helix III. Sie bedeuten jedoch keine direkte physische Interaktion [Hoffgaard et al., Prep].

NUGANNA

Durch die allgemeinere Beschreibung an den Positionen 3 und 8, in denen auch Wobble Basenpaare zugelassen wurden, gibt es rund 500000 zusätzliche Treffer, die dem Deskriptor (Abbildung 2.9(d)) entsprechen (siehe Abbildung 3.24). Am häufigsten wurde ein U3A8 Basenpaar gefunden, was ebenfalls von Perreault *et al.* beobachtet wurde [Perreault et al., 2011].

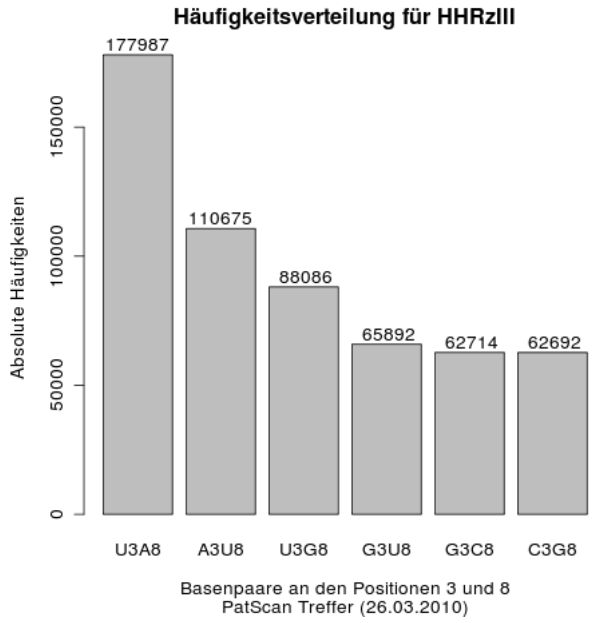


Abbildung 3.24: *Hammerhead* Ribozym Typ III N3N8
Die Treffer der PatScan Suche vom 26.03.2010 mit dem Deskriptor (Abbildung 2.9(d)) zeigen in den Positionen 3 und 8 des HHRz Typ III verschiedene Watson-Crick und Wobble Basenpaarungen. Die häufigste beobachtete Kombination ist ein U3A8 Basenpaar, was dem katalytischen Zentrum UH, UUGANAA, GAAA entspricht.

Vergrößerung der Loops

Eine weitere Modifizierung des HHRz-Motivs ist die Erweiterung der *Loops* mit mindestens 100 nt und maximal 1000 nt, ähnlich der Suche von Martick *et al.* in [Martick et al., 2008]. Das Ziel dieser Motivbeschreibung ist mögliche HHRz zu finden, die bestehend aus zwei Fragmenten, hunderte von Nukleotiden von einander getrennt in *trans* spalten können.

Die Suche vom 12.11.2010 ergab 7604 primäre Treffer und enthält alle von Martick *et al.* gefundenen Treffer, die dieser Beschreibung entsprechen. Nach Anwendung aller Filterschritte (Tabelle 2.3) verbleiben 274 HHRz III mit erweitertem *Loop* I. In dieser Menge fehlen jedoch einige der von Martick *et al.* identifizierten Treffer, da diese einen $\Delta\Delta G$ Wert größer als den gesetzten Schwellenwert von 0.5 kcal/mol besitzen (siehe Tabelle 3.6).

10

Tabelle 3.6: *Hammerhead* Ribozym Typ III Vergleich

Spezies	Position	Übereinstimmung mit Deskriptor 2.10(e)	$\Delta\Delta G$
Mus musculus	6:129136035-129136400:+	ja	$4.7\frac{kcal}{mol}$
Mus musculus	6:129042600-129043403:-	ja	$11.65\frac{kcal}{mol}$
Rattus norvegicus	4:166425123-166426033:+	ja	$2.87\frac{kcal}{mol}$

Tabelle 3.6 – Fortsetzung

Spezies	Position	Übereinstimmung mit Motiv 2.10(e)	$\Delta\Delta G$
Rattus norvegicus	4:166290554-166291351:-	ja	$4.47\frac{kcal}{mol}$
Rattus norvegicus	4:166331348-166332163:-	ja	$6.53\frac{kcal}{mol}$
Tupaia belangeri	scaffold_9056:3264:4881:-	nein, da L1=1500	-
Erinaceus europaeus	scaffold_303363:11370:12491:+	nein, da L1=1005	-
Equus caballus	6:36560178:36560961:+	nein, da Wobble Basenpaar	-
Loxodonta africana	scaffold_60948:6692:8548:-	nein, da L1=1739	-
Bos taurus	5:99619543:99621056:+	nein, da L1=1398	-
Canis familiaris	27:39360384:39361158:-	nein, da Wobble Basenpaar	-
Ornithorhynchus anatinus	Contig10595:5483:5745:-	nein, da Mismatch	-

[10] 3.6 zeigt die Ursache der Filterung der Treffer aus *Supplementary Figure* 2 [Martick et al., 2008]. *Hammerhead* Ribozyme deren Sequenzen mit dem Deskriptor 2.10(e) übereinstimmen, besitzen einen berechneten $\Delta\Delta G$ Wert.

Wird statt dessen der $\Delta\Delta G$ Wert der Suche vom 12.11.2010 auf 5 kcal/mol erhöht, bleiben nach der Filterung (Tabelle 2.3) von den ursprünglich 7604 Treffern 806 HHRz III Treffer mit einem erweiterten *Loop* I.

Anschließend wurden aus der Ausgangsmenge (7604 Treffer) 2197 redundante Sequenzen herausgefiltert und mit Hilfe des Skriptes getSupportingEvidence ausgewertet. Wie bereits beschrieben ist es nicht möglich, Annotationen für Sequenzen aus anderen Quellen als Ensembl und Ensemblgenomes auf diese Weise zu bestimmen. Deshalb wurden weitere 1102 Treffer entfernt. Insgesamt wurden 4305 *Slices* untersucht, wobei die Länge des *Slices* der Trefferposition entspricht.

Davon werden 1411 (33%) dieser *Loci* mit Genen assoziiert, von denen 356 (25%) in Exons liegen. Das heißt die anderen 1055 (75%) Treffer befinden sich in Introns. Repetetive Regionen existieren sowohl in intergenischen als auch in proteinkodierenden Regionen. 3312 (77%) der 4305 untersuchten Bereiche befinden sich in annotierten *Repeats*.

Wird statt *Loop* I *Loop* II mit mindestens 100 nt und maximal 1000 nt erweitert (Suche vom 03.12.2010), werden 6096 primäre Treffer gefunden, von denen 42 einzigartige Treffer nach der Filterung verbleiben. Die Auswertung der Lokalisation zeigt, dass nach dem Entfernen von 1180 redundanten Treffern, von 4126 untersuchten *Slices*, 1311 (32%) der *Slices* sich in Genen befinden, davon 333 (25%) in Exons und 978 (75%) in Introns. Die anderen 2815 (68%) Treffer befinden sich in intergenischen Bereichen. Ein Großteil der untersuchten Regionen 3199 (78%) ist repetitiv.

Die 11fach wiederholte Suche in Zufallssequenzen der Länge 271178200 Basenpaare mit gleichverteilter Nukleotidhäufigkeit ergibt für beide *Loop*-Erweiterungen (L1, L2) die annähernd gleiche Treffermenge mit durchschnittlich (16/11) und (15/11) Treffern. Dies entspricht für den DNA-Doppelstrang einer mittleren Motivwahrscheinlichkeit von rund $1.04 * 10^{-8}$. Die Menge der gefundenen primären Treffer in einem Suchraum der Größe $1.58 * 10^{11}$ Basenpaare ist somit 2 mal größer als die 3287 erwarteten Treffer in Zufallssequenzen.

Typ I

Die Tabelle 3.7 zeigt eine Übersicht der Suchergebnisse nach Typ I HHRz. Die Suchparameter befinden sich im Abschnitt 2.4.17.

[11]

Tabelle 3.7: *Hammerhead* Ribozym Typ I Ergebnisse

Datum	Suche	Faltung	Spezies	Primär	Filter
03. Sep. 2009	PatScan	keine	0V 0A 0B 3E	0V 0A 0B 48822 (3E)	-
01. Okt. 2009	RNAbob	UNAfold	0V 0A 0B 3E	0V 0A 0B 77723 (3E)	0V 0A 0B 10264 (3E)

Tabelle 3.7 – Fortsetzung

Datum	Suche	Faltung	Spezies	Primär	Filter
22. Mär. 2010	RNAbob	keine	796V 4A 442B 137E	1450 (109V) 1032 (4A) 56586 (397B) 2002236 (132E)	-
20. Dez. 2010	PatScan	Mfold	0V 0A 0B 3E	0V 0A 0B 66187 (2E)	0V 0A 0B 97 (2E)
07. Jan. 2011	PatScan	Mfold	3048V 50A 1227B 193E	2383 (33V) 837 (21A) 29696 (664B) 2346196 (178E)	1 (1V) 5 (4A) 85 (60B) 4512 (135E)

[11] Dargestellt sind die Ergebnisse der Typ I *Hammerhead* Ribozym Suchen mit den jeweiligen Such- und Faltungsprogrammen, der Anzahl durchsuchter Spezies (graue Balken der „Spezies“ Spalte), der Anzahl der primären Treffer des Suchprogrammes (graue Balken der „Primär“ Spalte) sowie der Anzahl der gefilterten Treffer (graue Balken der „Filter“ Spalte), jeweils mit der Anzahl der Spezies, eingeteilt in Viroide, Satelliten RNA, Viren (V), Archaeen (A), Bakterien (B) und Eukaryonten (E), in Klammern.

Eine erste Suche nach HHRz vom Typ I (03.09.2009) ergab in den Schistosomen 48822 primäre Treffer. Bei Verwendung von RNAbob und dem Zulassen von Wobble Basenpaaren wurden 77723 primäre Treffer gefunden, von denen nach der Faltung mit UNAfold und Filterung 10264 mögliche HHRz I verblieben (01.10.2009). Eine anschließende genomweite Suche mit RNAbob ermittelte 2061304 primäre Treffer.

Durch das Erstellen eines MSA mit weiteren HHRz I Beispielen, wie in Abschnitt 2.4.17 beschrieben, wurde der Deskriptor modifiziert. Die neue Suche mit PatScan ergab in den Schistosomen (*Schistosoma japonicum* und *Schistosoma mansoni*) 66187 primäre Treffer (20.12.2010). Die darauffolgende genomweite Suche (07.01.2011) mit PatScan und einem verallgemeinerten Deskriptor sowie mit N3N8 und erlaubten Wobble Basenpaaren ergab 2379112 primäre Treffer in 896 (20%) Genomen von denen 664 Genome signifikant mehr bzw. weniger Treffer besitzen als durch Zufall erwartet (Abbildung 3.25). Zur Abschätzung der HHRz I Motivwahrscheinlichkeit wurden mit dem Skript getObservedMotifProbability mehrere Zufallssequenzen erzeugt, nach dem Motiv durchsucht und das Ergebnis als Mittelwert zur Erzeugung eines linearen Modells verwendet und in die Abbildung 3.25 eingezeichnet. Nach Filterung aller primären Treffer, die eine größere freie Energie als -7 kcal/mol besitzen bleiben 2139166 Treffer.

Nach Anwendung der weiteren Filterschritte (Tabelle 2.3) verbleiben 4603 einzigartige HHRz I Kandidaten aus 200 Genomen (Abbildung 3.26). Die meisten sind in *Schistosoma mansoni* zu finden.

Die Analyse der Längenverteilung (Abbildung 3.27) zeigt, dass ein Großteil der HHRz I Kandidaten eine Länge zwischen 50 bis 60 nt besitzt. Die *Loop*-Größen variieren hauptsächlich zwischen 5, 7 oder 9 nt[Hoffgaard et al., Prep], was energetisch bevorzugte Größen für Kissing Komplexe sind [Tinoco & Bustamante, 1999]. Des Weiteren existieren auch bei diesen Sequenzen signifikante co-evolutionäre Zusammenhänge innerhalb und zwischen den Helices [Hoffgaard et al., Prep]. 2304 (50%) der 4603 HHRz I Treffer stammen aus Ensembl und Ensemblgenomes und wurden mit dem Skript getSupportingEvidence näher untersucht. Insgesamt gibt es 21159 Annotationen für diese *Slices*. 656 (28%) der Regionen werden mit Genen assoziiert, von denen 53 (8%) in Exons liegen. Die anderen 603 (92%) Treffer befinden sich in Introns. Der Großteil der 1648 (72%) Treffer kommt in intergenischen Bereichen vor. 950 (41%) der 2304 untersuchten *Slices* befinden sich in annotierten *Repeats*. Diese Ergebnisse beinhalten Annotationen aus beiden Strängen. Werden alle Annotationen des Treffer-Strangs ausgewählt, verbleiben 9087 Annotationen, bestehend aus 627 Genen, 36 Exons, 895 *Repeats* und weiteren Annotationen.

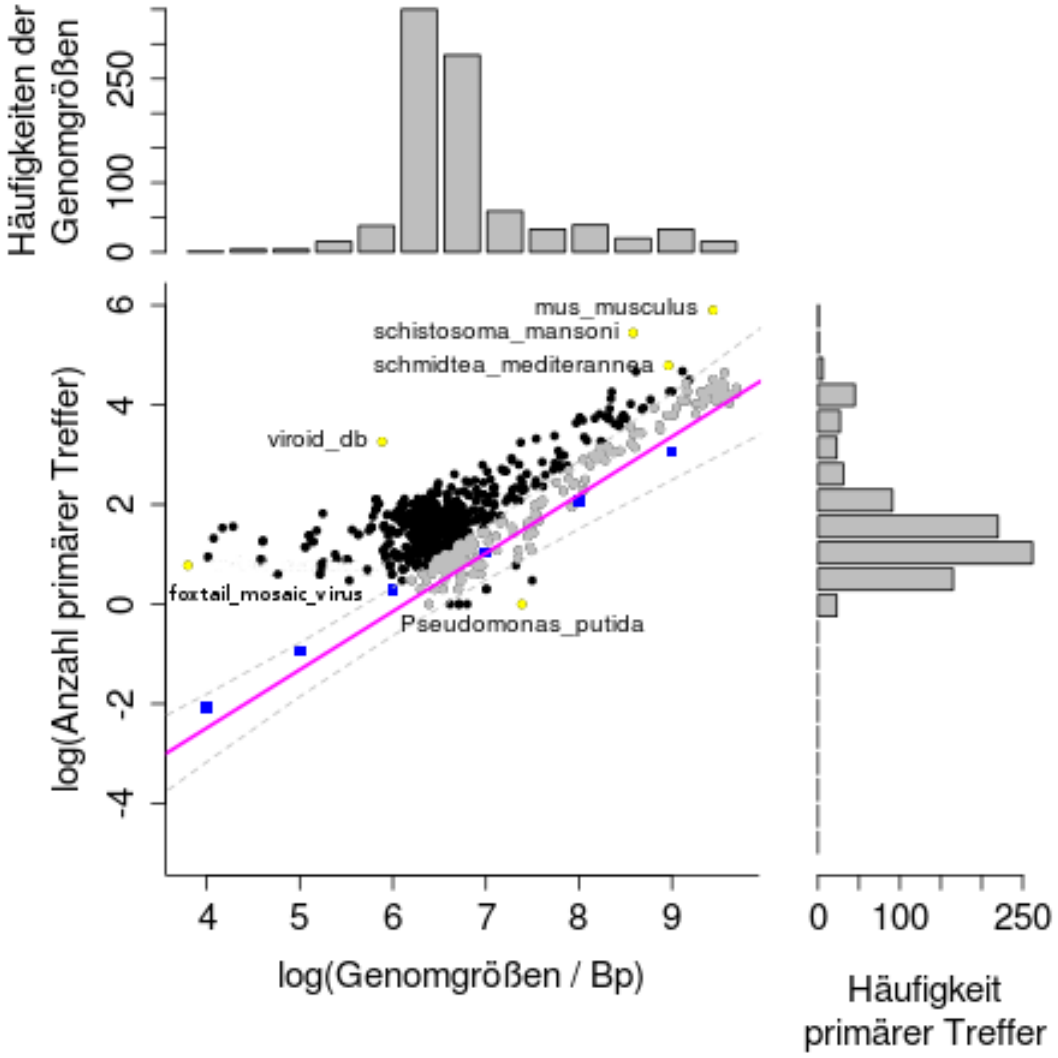


Abbildung 3.25: Treffer-Größenvergleich für Typ I *Hammerhead* Ribozyme
Die Daten sind zur besseren Darstellung logarithmisch aufgetragen und zeigen die Anzahl primärer Treffer (07.01.2011) unterschiedlicher Organismen mit deren jeweiligen Genomgrößen modifiziert aus [Hoffgaard et al., Prep]. Zusätzlich sind die Mittelwerte mehrerer Ziehungen der in Zufallssequenzen ermittelten Anzahl primärer Treffer in den jeweiligen Genomgrößen als blaue Quadrate eingezeichnet. Diese wurden zur Erstellung eines linearen Models verwendet (magenta). Mit diesem Model wurde ein 0.95 Konfidenzintervall ermittelt und als gestrichelte Linien eingezeichnet. Alle darin enthaltenen grauen Punkte entsprechen den erwarteten Treffergrößen, die durch Zufall entstanden sein könnten. Alle schwarzen Punkte sind signifikant größer oder kleiner und entsprechen nicht dem Zufall. Gelbe Punkte besitzen eine Beschriftung, sind besonders unter- oder überrepräsentiert und wurden zum besseren Vergleich mit 3.26 hervorgehoben.

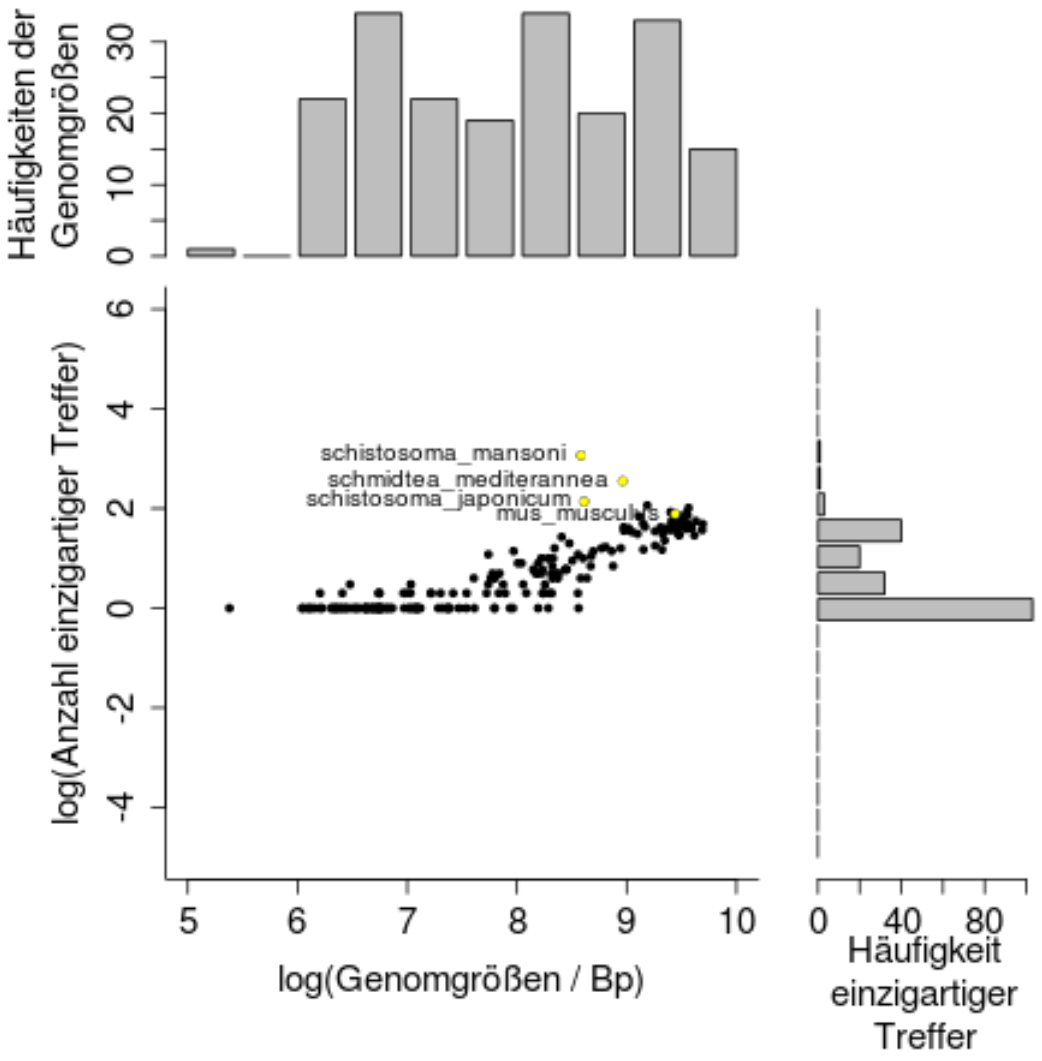


Abbildung 3.26: Einzigartige Treffer-Größenvergleich für Typ I *Hammerhead* Ribozyme

Ähnlich zu 3.25 sind logarithmisch die gefilterten Treffer nach Anwendung der *Pipeline* mit den Faltungs- und Filterparametern aus Tabelle 2.2 und 2.3 (07.01.2011) gegen die jeweiligen Genomgrößen aufgetragen. In *Schistosoma mansoni* befinden sich relativ betrachtet die meisten Treffer. Gelbe Punkte entsprechen den verbliebenen überrepräsentierten Treffern im Vergleich zu 3.25.

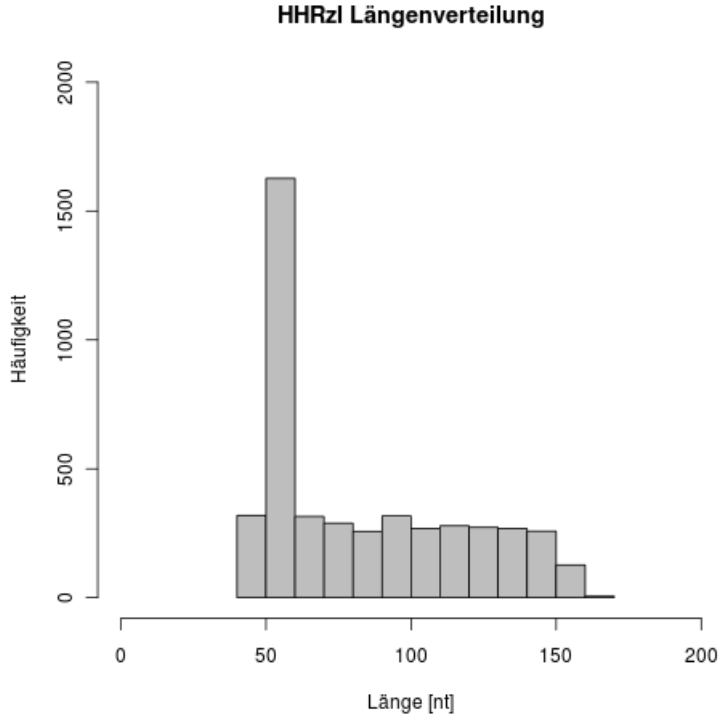


Abbildung 3.27: *Hammerhead* Ribozyme Typ I Längenverteilung
Zu sehen sind die unterschiedlichen HHRz I Längen der neu gefundenen Kandidaten.

Typ II

Tabelle 3.8 zeigt eine Übersicht der Suchergebnisse nach Typ II noch bevor Jimenez *et al.* und Perreault *et al.* ihre Ergebnisse präsentierten [Jimenez et al., 2011, Perreault et al., 2011]. Die Suchparameter befinden sich im Abschnitt 2.4.17.

[12]

Tabelle 3.8: *Hammerhead* Ribozym Typ II Ergebnisse

Datum	Suche	Faltung	Spezies	Primär	Filter
03. Sep. 2009	PatScan	keine	0V 0A 0B 3E	0V 0A 0B 459 (2E)	-
14. Sep. 2009	RNAbob	UNAfold	0V 0A 0B *Scma* (1E)	0V 0A 0B 5724 (1E)	0V 0A 0B 216 (1E)
16. Sep. 2009	PatScan	UNAfold	0V 0A 0B *Scja* (1E)	0V 0A 0B 5845 (1E)	0V 0A 0B 185 (1E)
28. Okt. 2009	RNAbob	Mfold	0V 0A 0B *Scma* (1E)	0V 0A 0B 5724 (1E)	0V 0A 0B 111 (1E)
02. Dez. 2009	RNAbob	Mfold	0V 0A 0B *Scma* (1E)	0V 0A 0B 5724 (1E)	0V 0A 0B 90 (1E)

Tabelle 3.8 – Fortsetzung

Datum	Suche	Faltung	Spezies	Primär	Filter
14. Jan. 2010	RNAbob	Mfold	0V 0A 0B *Scma* (1E)	0V 0A 0B 2004 (1E)	0V 0A 0B 51 (1E)
19. Mär. 2010	RNAbob	keine	531V 3A 348B 111E	368 (42V) 336 (3A) 16297 (261B) 532261 (104E)	-
03. Nov. 2010	PatScan	Mfold	0V 0A 0B *Xetr* (1E)	0V 0A 0B 416 (1E)	0V 0A 0B 0E

[12] Die Suchen nach Typ II *Hammerhead* Ribozymen mit den jeweiligen Such- und Faltungsprogrammen, der Anzahl durchsuchter Spezies (graue Balken der „Spezies“ Spalte), der Anzahl primärer Treffer des Suchprogrammes pro Gruppe (graue Balken der „Primär“ Spalte) sowie der Anzahl der gefilterten Treffer pro Gruppe (graue Balken der „Filter“ Spalte), jeweils mit der Anzahl der Spezies (Viroide, Satelliten RNA, Viren (V), Archaeen (A), Bakterien (B), Eukaryonten (E)) in Klammern werden in 3.8 zusammengefasst. Suchen in denen nur ein Organismus durchsucht wurde sind mit (Scja) *Schistosoma japonicum*, (Scma) *Schistosoma mansoni* und (Xetr) *Xenopus Tropicalis* abgekürzt.

Die erste Suche nach Typ II (03.09.2009) in den Schistosomen mit PatScan und Watson-Crick Basenpaarungsregeln ergab in *Schistosoma japonicum* 239 und in *Schistosoma mansoni* 220 mögliche HHRz mit einer passenden *Pattern*-Übereinstimmung. Werden zusätzlich Wobble Basenpaarungen erlaubt, können in *Schistosoma japonicum* weitere 5606 Treffer gefunden werden (16.09.2009). Bei Verwendung von RNAbob sind es in *Schistosoma mansoni* 5504 zusätzliche Treffer (14.09.2009). Nach der Faltung mit UNAfold und Anwendung der Filterschritte verbleiben 185 bzw. 216 potenzielle HHRz II. Die Faltung mit Mfold und Anwendung der gleichen Filterschritte führte zu 111 potenziellen HHRz II (28.10.2009). Durch das Hinzufügen eigener Faltungseinschränkungen wurden weitere 21 Treffer gefiltert (02.12.2009).

Durch die Einschränkung des *Loop* III auf NN[7] (14.01.2010) reduzierte sich die Treffermenge in *Schistosoma mansoni* um 35%. Dadurch wurde jedoch der beste Treffer mit dem kleinsten $\Delta\Delta G$ Wert entfernt.

Eine erste genomweite Suche nach Typ II HHRz in 993 genomischen Sequenzen (19.03.2010) führte zu 549262 primären Treffern.

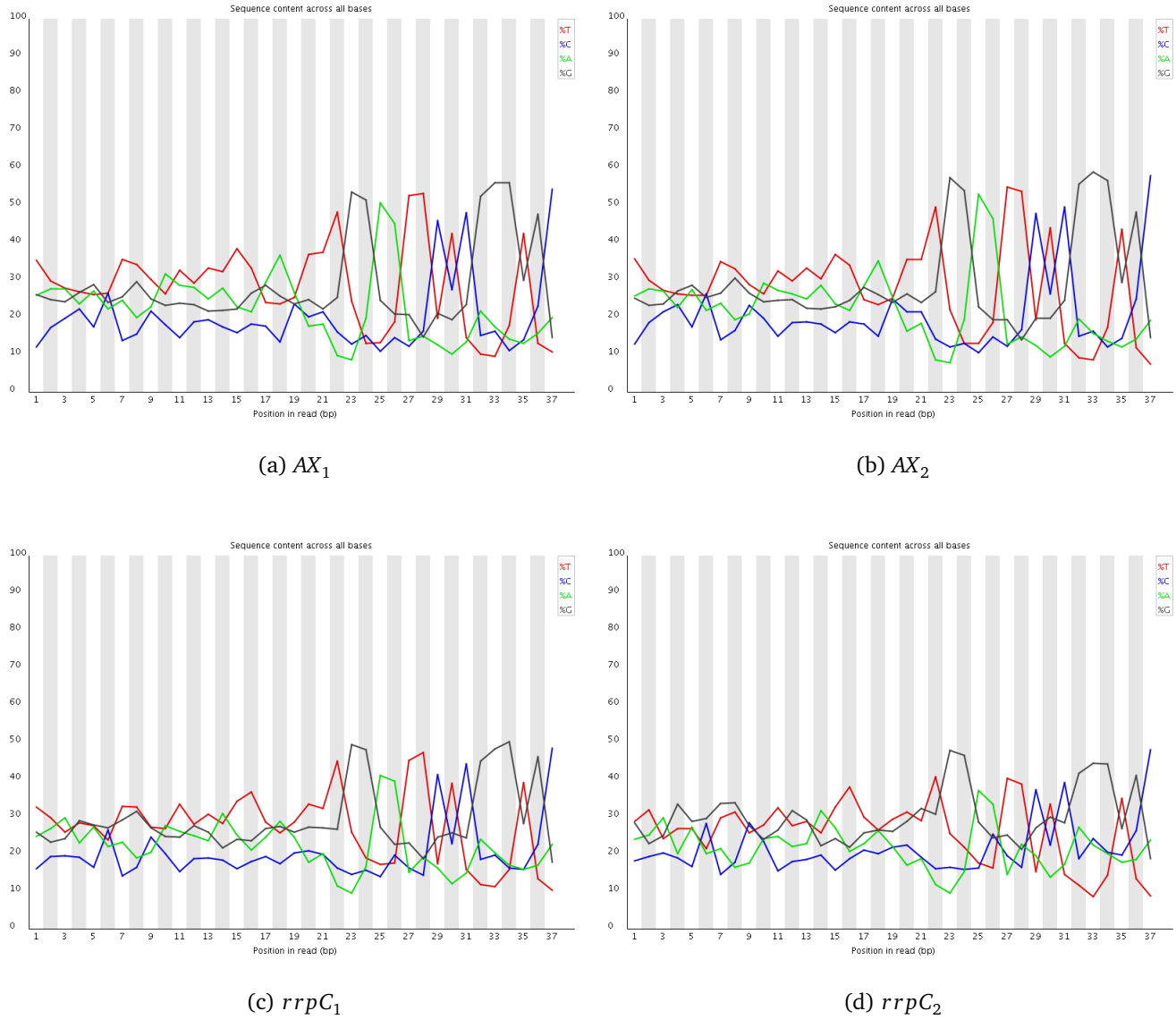
Eine wiederholte Suche in *Xenopus tropicalis* mit 416 primären Treffern und anschließender Anwendung aktueller Filtereinstellungen ergab keine Typ II Treffer.

Experimentell von Preeti Bajaj und im Rahmen dieser Arbeit getestete Sequenzen aus *Schistosoma mansoni* besaßen keine Aktivität (Daten nicht gezeigt).

3.3 Deep Sequencing

Auf den folgenden Seiten werden die Ergebnisse des zweiten Projektes präsentiert. Das Ziel war es herauszufinden, ob und wie *RNA-dependent RNA Polymerase*, speziell *RNA-directed RNA Polymerase C*, an der Regulierung von Retrotransposons und am RNA-Interferenz Mechanismus beteiligt sind.

Insgesamt gibt es 27 - 28 Millionen *Reads* für den Wildtyp (AX_1, AX_2) und 16 Millionen *Reads* für die zwei Replikate des *rrpC* Gendeletionsstamms ($rrpC_1$, $rrpC_2$). Dies zeigt bereits eine Reduzierung kleiner RNA im Gendeletionsstamm um 42%. Zu Beginn haben alle *Reads* eine Länge von 37 nt. Mehr als 92% der *Reads* enthalten Teile einer während der Sequenzierung verwendeten Adaptersequenz (siehe Abschnitt 2.4.18). Dies ist in Abbildung 3.28 zu erkennen, welche die Nukleotidhäufigkeit pro *Read*-Position zeigt.

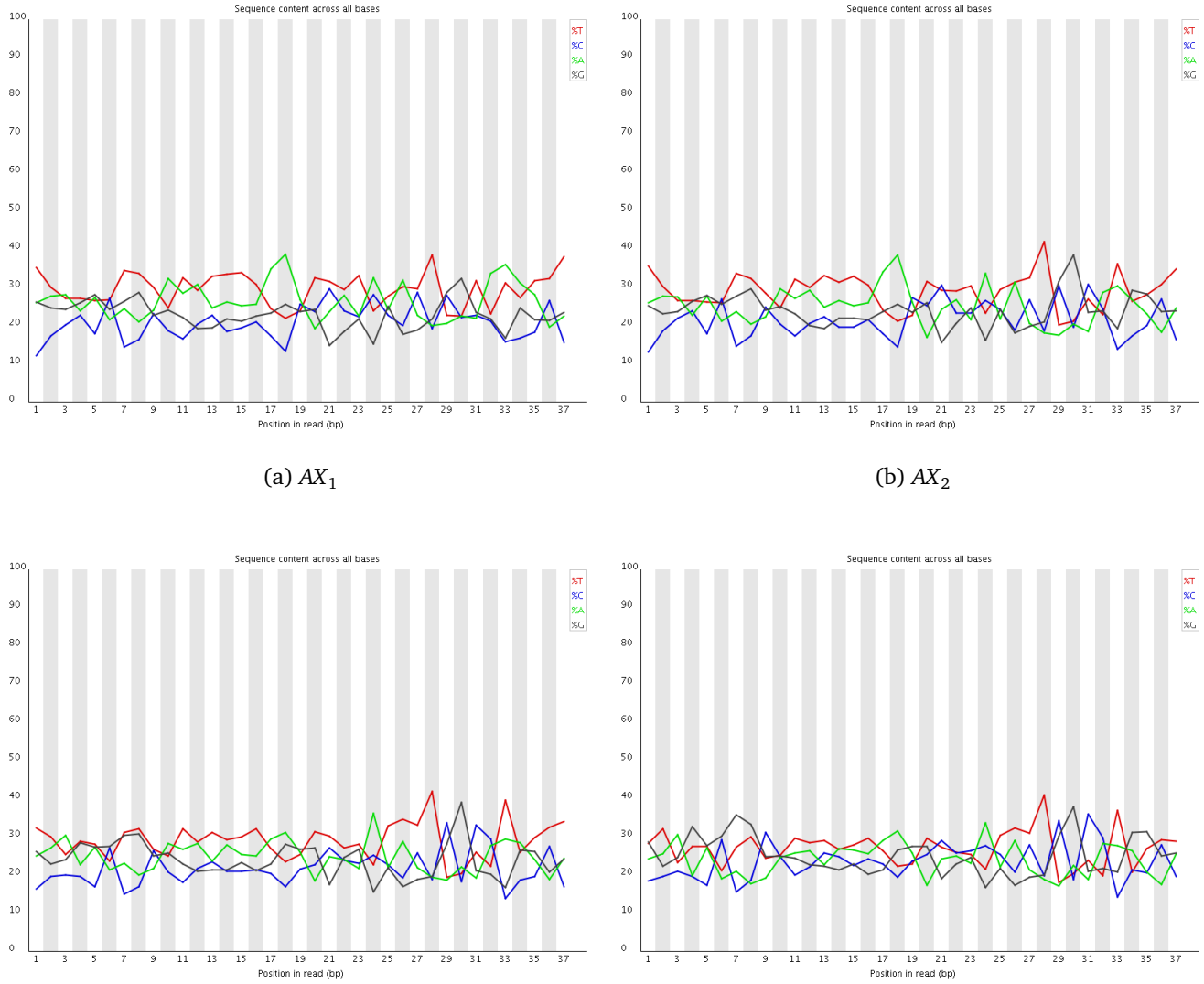


(a) AX_1 (b) AX_2

(c) $rrpC_1$ (d) $rrpC_2$

Abbildung 3.28: *Read* Nukleotidhäufigkeitsverteilung
Zu sehen sind die Nukleotidhäufigkeitsverteilungen der *Reads*, der einzelnen Replikate. Die x-Achse entspricht der jeweiligen *Read*-Position von 1 - 37 nt und die y-Achse dem prozentualen Anteil der Basen, Thymin in rot, Cytosin in blau, Adenin in grün und Guanin in schwarz. Beginnend mit Position 22 sind bestimmte Nukleotide stark überrepräsentiert.

Es gibt beginnend mit Position 22 ein Ungleichgewicht der Verteilung mit Unterschieden von > 20%, was für eine überrepräsentierte Sequenz spricht. In einer gleichverteilten Zufallssequenz wären minimale bis keine Unterschiede zu erwarten. Aus diesem Grund wurde im nächsten Schritt die Adaptersequenz entfernt. Daraus ergibt sich eine geglättete Nukleotidverteilung (Abbildung 3.29) und eine unterschiedliche Längenverteilung der *Reads* (siehe Abbildung 3.30), wobei eine Länge von 21 nt in allen Replikaten am häufigsten vorkommt.



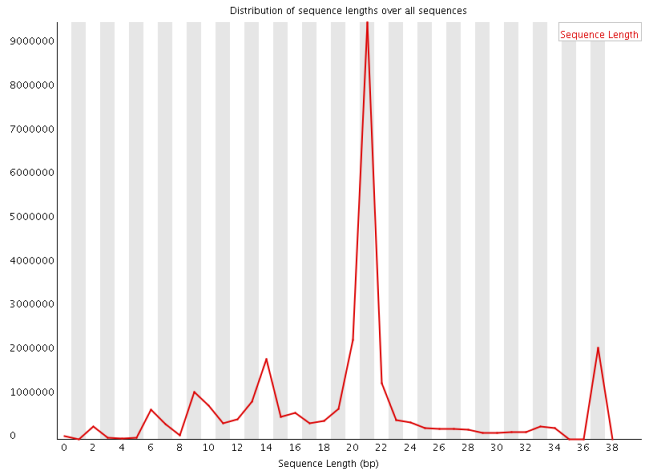
(a) AX_1 (b) AX_2

(c) $rrpC_1$ (d) $rrpC_2$

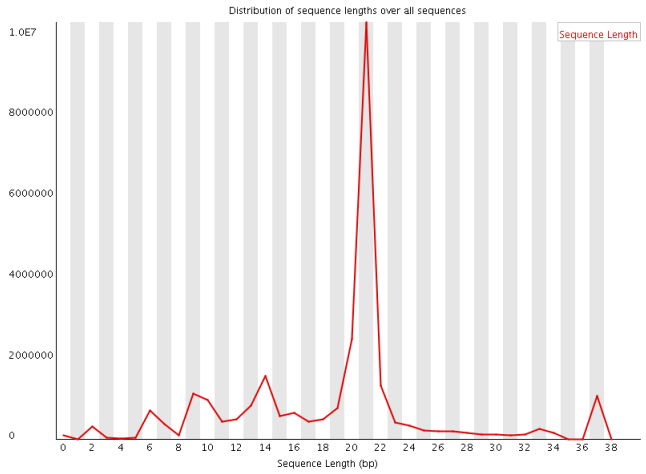
Abbildung 3.29: Gekürzte *Read* Nukleotidhäufigkeitsverteilung

Nach Entfernung der Adaptersequenz ergeben sich die obigen Nukleotidhäufigkeitsverteilungen der *Reads*, der einzelnen Replikate. Die x-Achse entspricht den *Read*-Position von 1 - 37 nt und die y-Achse dem prozentualen Anteil der Basen, Thymin in rot, Cytosin in blau, Adenin in grün und Guanin in schwarz.

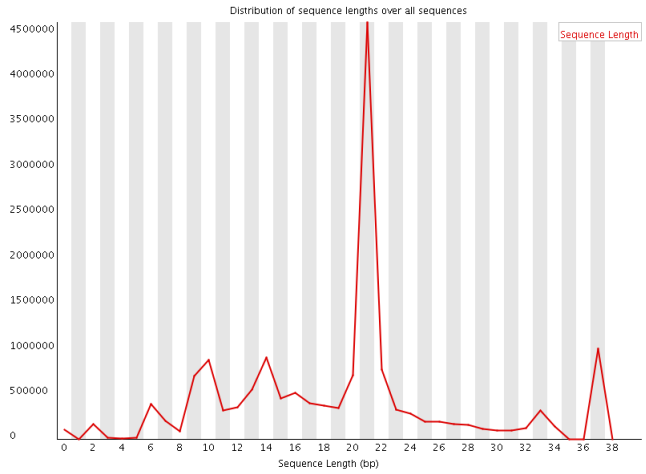
Außerdem ist im $rrpC_2$ Gendeletionsstamm ein Peak bei einer Länge von 10 nt zu sehen, der 8% der Ausgangsmenge entspricht.



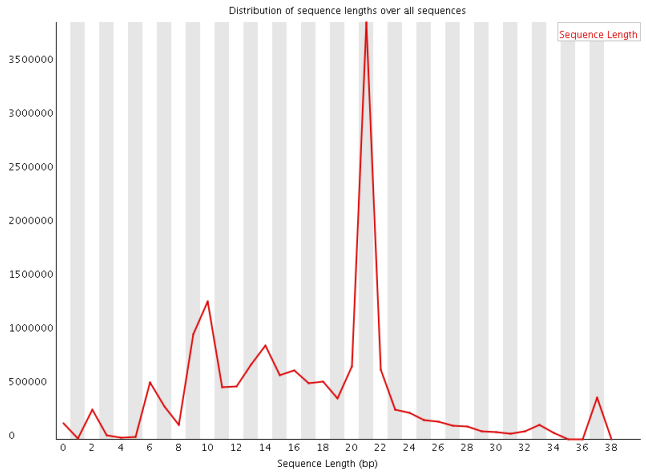
(a) AX_1



(b) AX_2



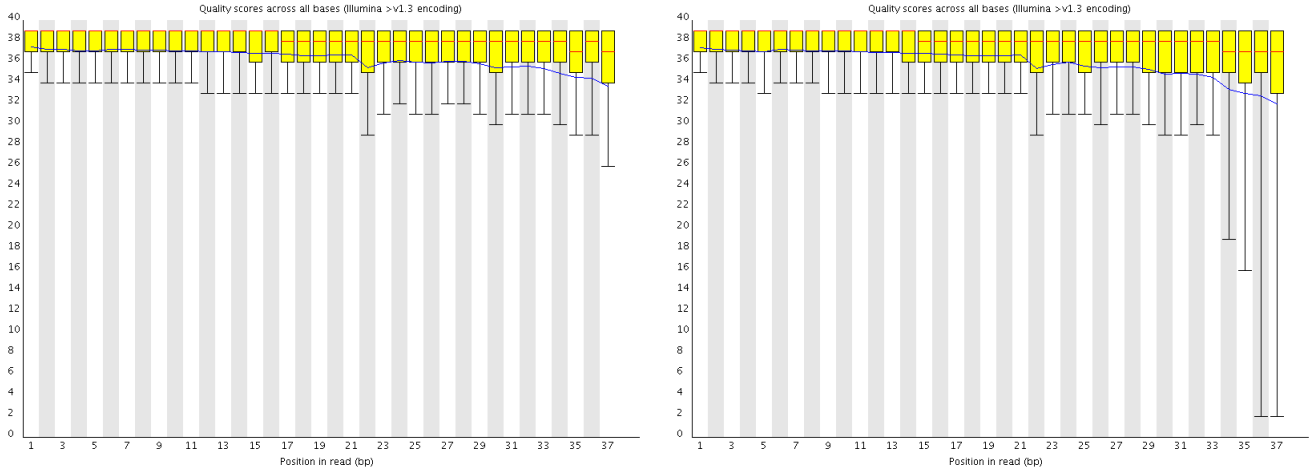
(c) $rrpC_1$



(d) $rrpC_2$

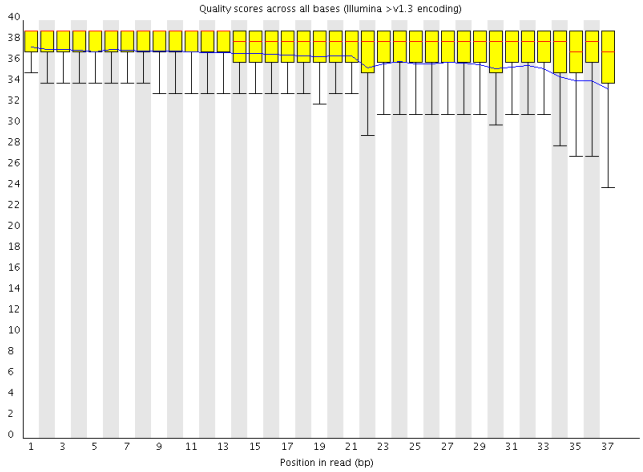
Abbildung 3.30: *Read*-Längenverteilung
Die Längenverteilung der *Reads* nach dem Entfernen der Adaptersequenz zeigt, dass in allen Replikaten der Großteil der *Reads* eine Länge von 21 nt besitzt. Auffällig ist ein zusätzlicher Peak in $rrpC_2$ mit mehr als 1000000 *Reads* (8%) mit einer Länge von 10 nt.

Die Qualität der Sequenzierungsergebnisse ist in Abbildung 3.31 zu sehen. Der Phread Quality Score pro *Read*-Position ist durchschnittlich größer als 30, was einer geschätzten Fehlerrate von 0,1%, also einem Fehler auf 10000 Basen entspricht und somit für eine hohe Qualität der *Reads* steht. Des Weiteren gibt es zahlreiche Duplikationen unter den *Reads*. Weniger als 5% sind einzigartig.

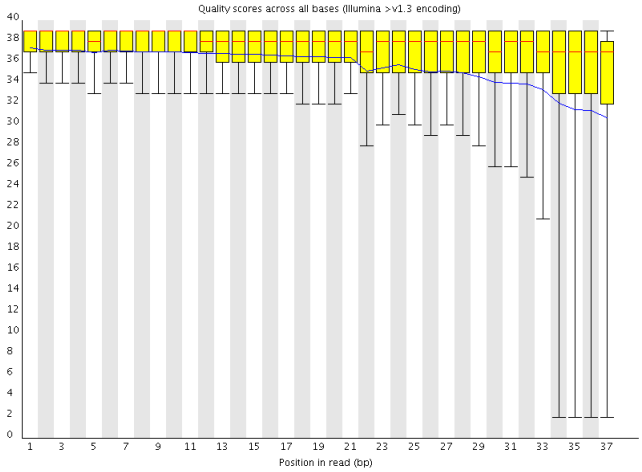


(a) AX_1

(b) AX_2



(c) $rrpC_1$



(d) $rrpC_2$

Abbildung 3.31: *Read Phred Quality Score*
Phread Quality Score der einzelnen *Read*-Positionen in Form eines Boxplots. Der durchschnittliche Score (blaue Linie) liegt über 30, was für eine hohe Qualität der *Reads* spricht.

Tabelle 3.9 zeigt die 10 häufigsten *Reads* nach dem Entfernen der Adaptersequenz.

[13]

Tabelle 3.9: Überrepräsentierte Sequenzen

AX_1		AX_2		$rrpC_1$		$rrpC_2$	
GGGTGCTGTATACT	5.34%	GGGTGCTGTATACT	4.00%	GTATACGGCC	3.92%	GTATACGGCC	5.24%
AGAGTGGTC	3.27%	AGAGTGGTC	3.27%	GGGTGCTGTATACT	3.64%	AGAGTGGTC	4.48%
GTGGTC	2.17%	GTGGTC	2.32%	AGAGTGGTC	3.34%	GTGGTC	2.86%
GTATACGGCC	1.78%	GTATACGGCC	2.21%	GTGGTC	2.06%	GGGTGCTGTATACT	2.70%
GGTGCTGTATACT	1.70%	GGTGCTGTATACT	1.39%	GGTGCTGTATACT	1.25%	CA	1.46%
CA	0.96%	CA	1.03%	TCCTTGTTGGTCTAGTGGTGAGGATTTTCGCCT	1.12%	GGTGCTGTATACT	1.26%
TAGGGTGCTGTATACT	0.88%	TAGGGTGCTGTATACT	0.80%	CA	0.94%	NULL	0.95%
AGGGTGCTGTATACT	0.71%	ATCCATTGAAGTTGTTAGATC	0.67%	TAGGGTGCTGTATACT	0.86%	GTATACGGCCA	0.93%
ATCCATTGAAGTTGTTAGATC	0.65%	AGGGTGCTGTATACT	0.64%	AGGGTGCTGTATACT	0.72%	TAGGGTGCTGTATACT	0.92%
TTAAGCATAAACGGTGAATACCTCGACTCCTAAATCT	0.63%	GTATACGGCCA	0.57%	NULL	0.66%	AGGGTGCTGTATACT	0.76%

[13] Die 10 häufigsten *Reads* unter den ersten 200000 Einträgen pro Replikat. Im Gendeletionsstamm gibt es zwei NULL-Einträge, was bedeutet, dass diese *Reads* komplett die Adaptersequenz enthielten und auf einen leeren String gekürzt wurden.

Im nächsten Schritt wurden die *Reads* gegen das *Dictyostelium discoideum* Genom aligniert (*Mapping*). Die zahlreichen Parametereinstellungen sind im Abschnitt 2.4.18 zu finden. Da *Dictyostelium discoideum* ein sehr AT reiches Genom mit vielen *Repeats* besitzt, wurden die *Reads* zusätzlich gegen eine ausgewählte Bibliothek von einfachen bis zu komplexen *Repeats* aligniert. Tabelle 3.10 zeigt einen Teil der *Mapping*-Ergebnisse.

[14]

Tabelle 3.10: *Mapping*-Ergebnis

Referenz	Parameter	AX_1	AX_2	$rrpC_1$	$rrpC_2$
		27909469 (100.00%)	28070029 (100.00%)	16262796 (100.00%)	16251369 (100.00%)
dicty_chromosomal	Standardeinstellung	91.81%	89.76%	90.27%	87.78%
1 mal aligned		1626098 (5.83%)	1231240 (4.39%)	1648990 (10.14%)	1283324 (7.90%)
>1 mal aligned		23997850 (85.98%)	23964328 (85.37%)	13031508 (80.13%)	12982743 (79.89%)
dicty_chromosomal	-M 100 -N 2 -L 20	96.41%	95.76%	94.63%	92.17%
1 mal aligned		1098628 (3.94%)	745345 (2.66%)	1163868 (7.16%)	877883 (5.40%)
>1 mal aligned		25808138 (92.47%)	26133896 (93.10%)	14225253 (87.47%)	14101224 (86.77%)
dicty_chromosomal	-N 0 -L 20 -R 3 -D 20 -i S 1, 0.50	97.30%	96.74%	95.39%	93.76%
1 mal aligned		1210816 (4.34%)	825356 (2.94%)	1299368 (7.99%)	962594 (5.92%)
>1 mal aligned		25944719 (92.96%)	26328770 (93.80%)	14214503 (87.41%)	14274108 (87.83%)

Tabelle 3.10 – Fortsetzung

Referenz	Parameter	AX_1	AX_2	$rrpC_1$	$rrpC_2$
dicty_repeats_j	-M 500 -N 1 -L 20 -R 3 -D 20 -i S,1,0.50	57.94%	62.82%	42.82%	44.87%
1 mal aligned		13214687 (47.35%)	14538930 (51.80%)	5030245 (30.93%)	4714539 (29.01%)
>1 mal aligned		2956012 (10.59%)	3095281 (11.03%)	1933180 (11.89%)	2576926 (15.86%)
dicty_repeats_j	-k 100	56.89%	61.62%	42.19%	44.38%
1 mal aligned		12925888 (46.31%)	14207014 (50.61%)	4935475 (30.35%)	4639974 (28.55%)
>1 mal aligned		2950474 (10.57%)	3091064 (11.01%)	1926406 (11.85%)	2572489 (15.83%)

[14] 3.10 zeigt einen Ausschnitt der *Mapping*-Ergebnisse gegen das *Dictyostelium discoideum* Genom (dicty_chromosomal) und die *Repeat*-Bibliothek (dicty_repeats_j) (siehe Abschnitt 2.4.18). Die dargestellte prozentuale *Alignment*-Rate setzt sich aus der Anzahl der *Reads*, die einmal oder mehrmals aligniert wurden, zusammen.

Die hohe Anzahl an Duplikationen ist mit mehr als 79% auch im *Mapping*-Ergebnis zu sehen, wobei mehr als 42% der *Reads* repetitiven Regionen zugeordnet wurden. Für die nachfolgenden Schritte wurde das *Mapping*-Ergebnis unter Verwendung der Standardeinstellung (dicty_chromosomal) und -M 500 -N 1 -L 20 -R 3 -D 20 -i S,1,0.50 (dicty_repeats_j) genutzt. Abbildung 3.32 zeigt die Verteilung der *Reads* aufgeschlüsselt in die einzelnen Chromosomen.

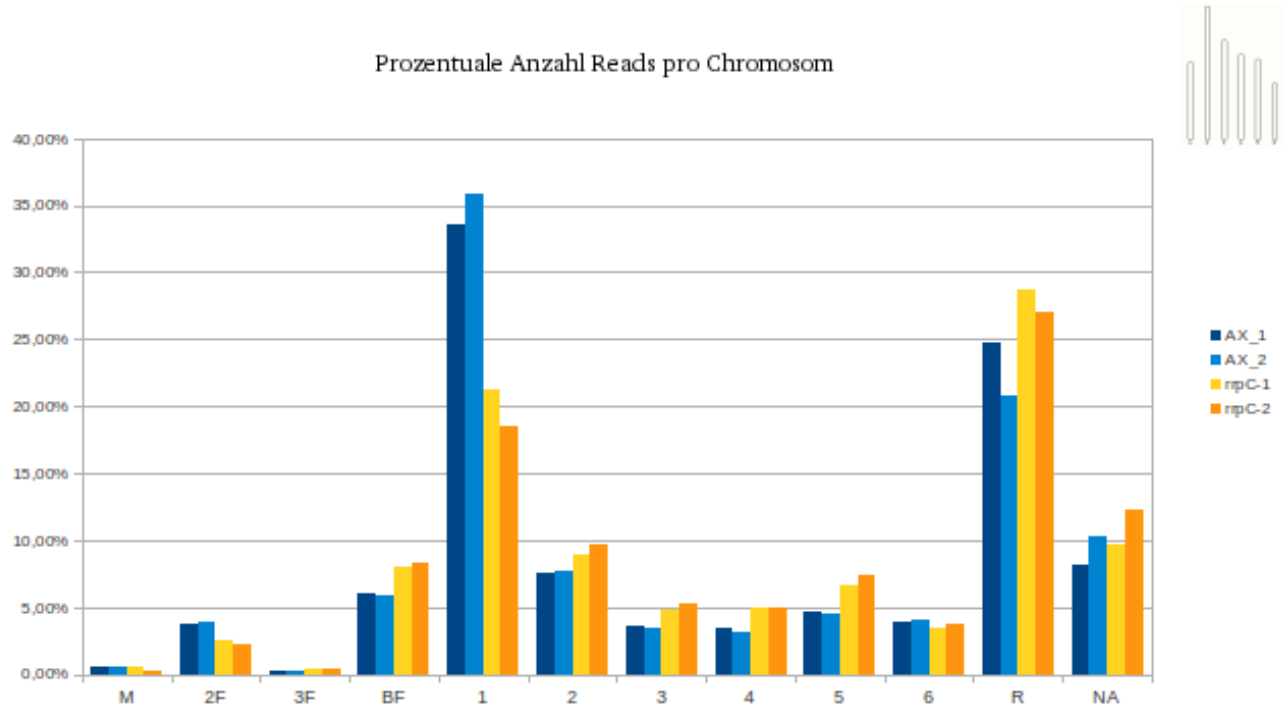


Abbildung 3.32: *Reads* pro Chromosom

3.32 zeigt den prozentualen Anteil der *Reads* pro Chromosom. In hell und dunkelblau ist der Wildtyp und in gelb und orange der *rrpC* Gendeletionsstamm dargestellt. NA entsprechen nicht alignierten *Reads*, 1-6 den Chromosomen, R dem extra-chromosomalen rDNA-Palindrom, 2F, 3F und BF nicht zuordbaren *Contigs* der Chromosomen 2-6 und M mitochondrialer DNA.

Im Wildtyp wurden die meisten *Reads* Chromosom 1 zugeordnet, gefolgt von dem extra-chromosomalen rDNA-Palindrom. Im Gendeletionsstamm ist es umgekehrt.

Tabelle 3.11 und Abbildung 3.33 zeigen eine Übersicht, wieviel der alignierten *Reads* mit den jeweiligen von dictybase gegebenen Annotationen überlappen. Der prozentuale Unterschied zwischen Wildtyp und Gendeletionsstamm ist minimal.

15

Tabelle 3.11: *Read*-Überlappungen zu annotierten Regionen

	AX_1		AX_2		$rrpC_1$		$rrpC_2$	
Chromosomen	25623948	91,811%	25195568	89,760%	14680498	90,270%	14266067	87,784%
Contigs	25623948	91,811%	25195568	89,760%	14680498	90,270%	14266067	87,784%
Intergenisch	12325604	44,163%	12861850	45,821%	6584935	40,491%	6623004	40,754%
Gene	13298344	47,648%	12333718	43,939%	8095563	49,780%	7643063	47,030%
Exons	13282339	47,591%	12326600	43,914%	8062830	49,578%	7614646	46,855%
Introns	16005	0,057%	7118	0,025%	32733	0,201%	28417	0,175%
Gaps	372925	1,336%	428921	1,528%	354830	2,182%	432164	2,659%
nicht aligniert	2285521	8,189%	2874461	10,240%	1582298	9,730%	1985302	12,216%
Repeats	16170699	57,940%	17634211	62,822%	6963425	42,818%	7291465	44,867%
Pseudogene	93779	0,336%	95270	0,339%	58158	0,358%	76298	0,469%
class_I_RNA	137877	0,494%	71453	0,255%	71461	0,439%	49757	0,306%
class_II_RNA	45981	0,165%	28441	0,101%	17712	0,109%	20022	0,123%
H_ACA_box snoRNA	2939	0,011%	2195	0,008%	1332	0,008%	1149	0,007%
C_D_box snoRNA	365386	1,309%	137250	0,489%	142174	0,874%	41662	0,256%
snRNA	17061	0,061%	12317	0,044%	11064	0,068%	9651	0,059%
tRNA	661958	2,372%	679016	2,419%	620211	3,814%	476553	2,932%
rRNA	6452491	23,119%	5434661	19,361%	4279338	26,314%	4039880	24,859%
ncRNA	251	0,001%	223	0,001%	230	0,001%	183	0,001%
RNase_P_RNA	1186	0,004%	759	0,003%	537	0,003%	431	0,003%
RNase_MRP RNA	1284	0,005%	981	0,003%	714	0,004%	554	0,003%
SRP_RNA	66286	0,238%	43017	0,153%	52879	0,325%	36431	0,224%

[15] Die Anzahl der *Reads*, die mit annotierten Positionen im Genom von *Dictyostelium discoideum* überlappen wurden mit HTseq ermittelt. Bei den Annotationen unterhalb der Trennlinie gibt es *Reads*, die verworfen wurden, da sie zu Chromosomen alignieren, die das Feature nicht enthalten. Ein Beispiel ist das extra-chromosomalen rDNA-Palindrom, dass keine annotierten *transfer RNA* (tRNA) enthält, weshalb *Reads*, die zum rDNA-Palindrom aligniert wurden, nicht in der Zählung der tRNA enthalten sind. Die Anzahl der Überlappungen mit *Repeats* stammt aus dem *Mapping*-Ergebnis (dict_repeats_j) mit Bowtie2.

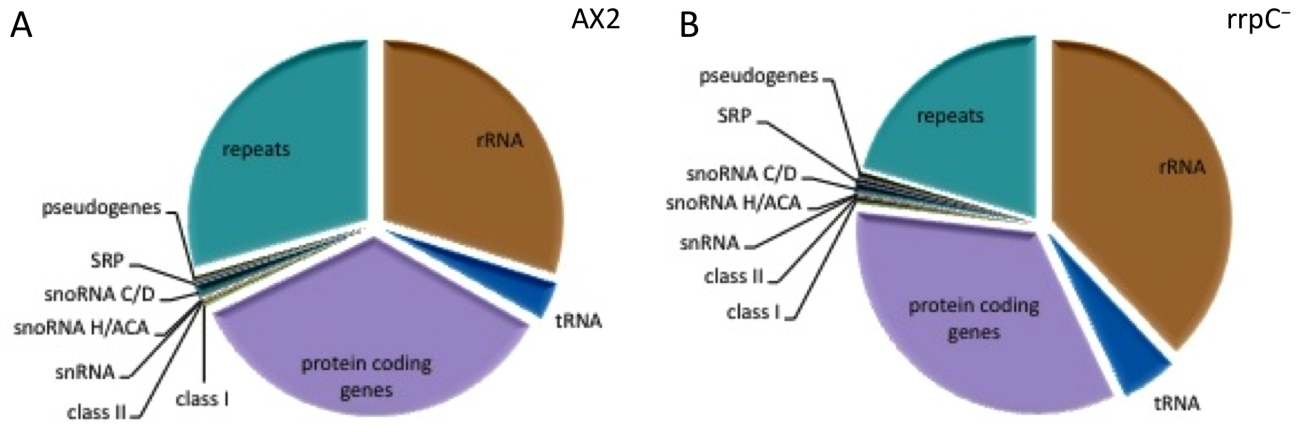


Abbildung 3.33: Diagramm der *Read*-Überlappungen zu annotierten Regionen
Die Kreisdiagramme fassen die Daten aus Tabelle 3.11 für den A) *rrpC* Wildtyp und B) *rrpC* Gendeletionsstamm zusammen. 3.33 wurde aus [Wiegand et al., Prep] übernommen und mit Gimp modifiziert.

Mehr als 43% der *Reads* alignieren zu annotierten Genen, von denen ein Großteil für ribosomale Bestandteile kodiert (siehe Tabelle 3.12).

[16]

Tabelle 3.12: Top 10 der häufigsten *Read*-Überlappungen mit Genen

Gen Produkt	Gen Name	AX_1	
	5S_rRNA-1	2062561	7,39%
	26S_rRNA-1	1438811	5,16%
	5S_rRNA-2	1221033	4,37%
	17S_rRNA-1	694489	2,49%
	5.8S_rRNA-1	463618	1,66%
DIRS1 ORF2 Fragment	*DDB_G0267236_RTE*	377392	1,35%
	26S_rRNA-2	253391	0,91%
DIRS1 ORF2 Fragment	*DDB_G0272554_RTE*	244605	0,88%
DIRS1 ORF2 Fragment	*DDB_G0267294_RTE*	241918	0,87%
DIRS1 ORF2 Fragment	*DDB_G0267210_RTE*	208357	0,75%
Gen Produkt	Gen Name	AX_2	
	5S_rRNA-1	1626913	5,80%
	26S_rRNA-1	1325541	4,72%
	5S_rRNA-2	1067684	3,80%
	17S_rRNA-1	639509	2,28%
DIRS1 ORF2 Fragment	*DDB_G0267236_RTE*	418261	1,49%
	5.8S_rRNA-1	346165	1,23%
DIRS1 ORF2 Fragment	*DDB_G0272554_RTE*	262741	0,94%
DIRS1 ORF2 Fragment	*DDB_G0267294_RTE*	260772	0,93%
DIRS1 ORF2 Fragment	*DDB_G0267210_RTE*	223963	0,80%
	26S_rRNA-2	223425	0,80%
Gen Produkt	Gen Name	$rrpC_1$	
	26S_rRNA-1	1342998	8,26%
	5S_rRNA-1	952275	5,86%
	5S_rRNA-2	713981	4,39%
	17S_rRNA-1	600663	3,69%
	5.8S_rRNA-1	273042	1,68%
	26S_rRNA-2	229472	1,41%
DIRS1 ORF2 Fragment	*DDB_G0267236_RTE*	118609	0,73%
DIRS1 ORF2 Fragment	*DDB_G0267294_RTE*	90026	0,55%
DIRS1 ORF2 Fragment	*DDB_G0272554_RTE*	87593	0,54%
	5.8S_rRNA-2	78564	0,48%
Gen Produkt	Gen Name	$rrpC_2$	
	26S_rRNA-1	1311817	8,07%
	5S_rRNA-1	847012	5,21%
	5S_rRNA-2	740017	4,55%
	17S_rRNA-1	609222	3,75%
	26S_rRNA-2	219272	1,35%
	5.8S_rRNA-1	206619	1,27%
DIRS1 ORF2 Fragment	*DDB_G0267236_RTE*	103377	0,64%
DIRS1 ORF2 Fragment	*DDB_G0267294_RTE*	73974	0,46%
DIRS1 ORF2 Fragment	*DDB_G0272554_RTE*	72111	0,44%
DIRS1 ORF2 Fragment	*DDB_G0267210_RTE*	61619	0,38%

[16] Unter Verwendung der dictybase Annotation sind die Top 10 Gene pro Replikat mit den meisten *Read*-Überlappungen dargestellt.

Die häufigsten *Read*-Überlappungen mit *Repeats* sind in allen Replikaten an erster Stelle mit DIRS-1, was frühere Studien bestätigt [Hinas et al., 2007], wobei die Menge im Gendeletionsstamm um rund 32% reduziert ist. Im Wildtyp folgt an zweiter Stelle ein AT *Repeat* (AX_1) und TRE5_B (AX_2). Im Gendeletionsstamm ist die Überlappung mit Skipper an zweiter Stelle.

Bei Betrachtung der *Read*-Verteilung, speziell bei DIRS-1 (Abbildung 3.34), fallen 3 Klassen kleiner RNA auf:

- solche, die in Wildtyp und Gendeletionsstamm annähernd konstant sind, wie z. B. in den LTR,
- solche die im Gendeletionsstamm halb oder mehr reduziert sind, wie z. B. am 3' Ende des ORF I bzw. 5' Ende des ORF II und
- solche, die nur im *rrpC*-Gendeletionsstamm auftreten, wie am 5' Ende des ORF I.

Abbildung 3.34: DIRS-1 *Read*-Verteilung

Mit dem Programm IGV kann das *Mapping*-Ergebnis des Retrotransposons DIRS-1 dargestellt werden. 3.34 besteht aus 6 Zeilen, die als „Tracks" bezeichnet werden. Track 1 zeigt den ausgewählten Bereich, welcher der gesamten Sequenz entspricht. Track 2 zeigt die dazu gehörige bekannte Annotation mit den linken und rechten LTR sowie den 3 möglichen ORF. Tracks 3 und 4 zeigen den Wildtyp AX_1 und AX_2 und Tracks 5 und 6 den Gendeletionsstamm $rrpC_1$ und $rrpC_2$. Pro Replikat sind die *Reads* beider Stränge in hellgrau und für den *Antisense* Strang in dunkelgrau dargestellt. Die Tracks sind entsprechend dem höchsten Peak in AX_2 Track 4 auf 396053 *Reads* skaliert. Die farbigen Balken entsprechen *Mismatches* an den jeweiligen Positionen und sind ab 20% *Mismatches* proportional in grün (Adenin), blau (Cytosin), braun (Guanin) und rot (Thymin) eingefärbt.

Die Abbildung 3.35 zeigt die *Read*-Verteilung für Skipper.

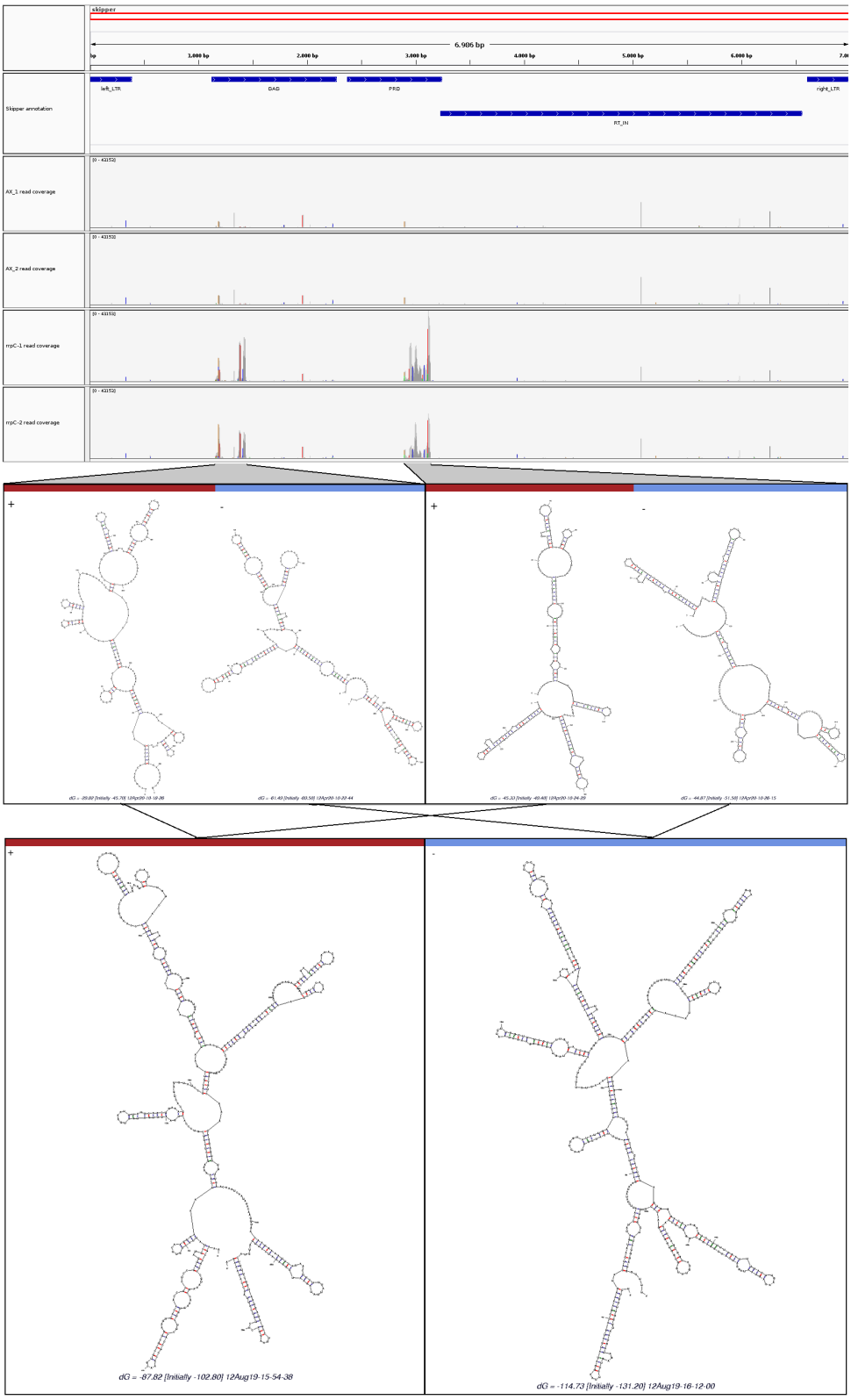


Abbildung 3.35: Skipper *Read*-Verteilung

Das *Mapping*-Ergebnis für das Retrotransposon Skipper zeigt in Track 1 den ausgewählten Bereich, in Track 2 die dazu gehörige Annotation, in den Tracks 3 und 4 den Wildtyp AX_1, AX_2 und in den Tracks 5 und 6 den Gendeletionsstamm $rrpC_1$, $rrpC_2$. Pro Replikat sind die *Reads* beider Stränge in hellgrau und für den *Antisense* Strang in dunkelgrau dargestellt. Die Tracks sind entsprechend dem höchsten Peak in $rrpC_1$ Track 5 auf 43153 *Reads* skaliert. Die farbigen Balken in den Tracks entsprechen *Mismatches* an den jeweiligen Positionen und sind ab 20% *Mismatches* proportional in grün (Adenin), blau (Cytosin), braun (Guanin) und rot (Thymin) eingefärbt. Zusätzlich sind die mit Mfold erstellten Faltungen des grau markierten Sequenzbereichs sowie deren Fusion dieser Bereiche, für den *Sense*- und *Antisense* Strang dargestellt.

Auffällig in der *Read*-Verteilung in Skipper sind zwei Regionen (1157-1443 entspricht dem 5' Ende von GAG und 2887-3138 dem 3' Ende von PRO), in denen im Gendeletionsstamm vermehrt kleine RNA auftreten. Die Mehrheit, der auf dem *Sense* Strang abgebildeten kleinen RNA, können ORF II zugeordnet werden. Die kleinen RNA des *Antisense* Strangs sind gleichmäßig auf die zwei Regionen verteilt. Die Ähnlichkeitssuche der zwei auffälligen Regionen mit BLASTN in einer miRNA Datenbank [Kozomara & Griffiths-Jones, 2011] (http://www.mirbase.org) ergab keine signifikanten Treffer. Auch die Vorhersage neuer miRNA mit dem Programm miR-abela konnte keine *pre*-miRNA identifizieren.

Beide Abbildungen (3.34 und 3.35) bestätigen die bisherigen experimentellen Daten von Stephan Wiegand, die im *rrpC*-Gendeletionsstamm eine verminderte Menge kleiner RNA für DIRS-1 und eine Anreicherung kleiner RNA für Skipper zeigen. Abbildung 3.36 zeigt die *Read*-Verteilung aller *Reads* größer einer Länge von 15nt beim extra-chromosomalen rDNA-Palindrom.

Abbildung 3.36: *Read*-Verteilung des extra-chromosomalen rDNA-Palindroms
Nach Auswahl aller *Reads* > 15nt wurden diese gegen das extra-chromosomale rDNA-Palindrom aligniert. Track 1 zeigt das gesamte Chromosom, Track 2 die dazu gehörige Annotation, Tracks 3 und 4 den Wildtyp AX_1, AX_2 und Tracks 5 und 6 den Gendeletionsstamm $rrpC_1$, $rrpC_2$. Zusätzlich ist in den Tracks 3-6 die Stranginformation farblich kodiert, wobei rot dem *Sense* Strang und blau dem *Antisense* Strang entspricht. Farbige Balken stehen für *Mismatches* an den jeweiligen Positionen und sind ab 20% *Mismatches* proportional in grün (Adenin), blau (Cytosin), braun (Guanin) und rot (Thymin) markiert.

Das *Mapping* mit 100 *Alignments* pro *Read* zeigt zwei Regionen. Zum einen auf dem *Antisense* Strang im 5' Bereich, wo sich die annotierte ribosomale RNA (rRNA) befindet [Boesler et al., 2011] und zum anderen auf dem *Sense* Strang im 3' Bereich, der dieser Region auf mehr als 7500 nt zu 99% gleicht. In Abbildung 3.36 befindet sich die höchste Anzahl kleiner RNA am 5' Ende des 5.8S rRNA Gens.

Die statistische Auswertung der Anzahl kleiner RNA im Wildtyp und Gendeletionsstamm zeigte, dass sich 2096 von 19233 untersuchten Regionen entlang des *Dictyostelium discoideum* Genoms signifikant unterscheiden, wobei im *rrpC* Gendeletionsstamm die Anzahl in 186 Regionen vermindert und in 1910 Regionen erhöht ist (Tabelle 3.13). Alle 19 untersuchten *Repeat*-Regionen unterscheiden sich ebenfalls signifikant, wobei die kleinen DIRS-1 RNA im Gendeletionsstamm vermindert und die Anzahl kleiner RNA in alle anderen *Repeats* erhöht sind (Tabelle 3.14).

[17]

Tabelle 3.13: Top 10 signifikant verminderte und erhöhte Mengen kleiner RNA pro *Locus*

Gen	Locus	Wert_1	Wert_2	log_2FoldChange	p-value
DDB_G0267272_RTE	DDB0232428:88914-90691	332,048	32,5411	-3,35106	0
DDB_G0269420	DDB0232428:2730079-2730479	189,416	13,6321	-3,79648	0
DDB_G0276981_RTE	DDB0232429:7435824-7437462	17,0063	1,65252	-3,36333	0
DDB_G0273423	DDB0232429:2975859-2977092	15,1766	1,36813	-3,47157	0
DDB_G0273835	DDB0232429:3465225-3466020	104,358	1,70549	-5,93521	0
fray2	DDB0232429:6994953-6998137	110,969	6,59196	-4,07331	0
tRNA-Cys-GCA-5	DDB0232430:1671337-1671409	79491,4	1950,07	-5,3492	0
DDB_G0279845	DDB0232430:2626208-2626828	102,663	2,19812	-5,5455	0
rps10	DDB0232431:4072415-4073194	3108,75	120,095	-4,69409	0
DDB_G0290223	DDB0232432:3816108-3817002	46,3964	5,60043	-3,0504	0
DDB_G0294204_RTE	DDB0215018:95618-95986	46,207	289,112	2,64545	0
DDB_G0294242_RTE	DDB0215018:122434-123171	6,4843	52,9844	3,03055	0
DDB_G0294164_RTE	DDB0215018:67861-68308	2,66027	61,9152	4,54065	0
DDB_G0294228_RTE	DDB0215018:110053-110728	0,612616	59,7937	6,60887	0
DDB_G0294232_RTE	DDB0215018:112861-113167	41,6281	263,479	2,66206	0
DDB_G0294176_RTE	DDB0215018:73570-74226	8,08137	49,0958	2,60293	0
DDB_G0294164_RTE	DDB0215018:66170-66673	17,042	157,39	3,20717	0
DDB_G0294368	DDB0220052:24815-25288	36,5253	827,326	4,50149	0
-	DDB0220052:20537-20662	1730,55	28723,6	4,05293	0
DDB_G0267592	DDB0232428:383615-385211	1,00496	8,77429	3,12615	0

[17] 3.13 zeigt die 10 größten log_2FoldChange der Anzahl der Fragmente pro *Locus* sortiert nach dem *P-Value*. Der FPKM Wert_1 entspricht dem Wildtyp und FPKM Wert_2 dem Gendeletionsstamm. Die Region DDB0220052:20537-20662 auf dem Chromosom BF, welches nicht zuordenbare *Contigs* aus den Chromosomen 4, 5 und 6 enthält, besitzt kein annotiertes Gen. Der log_2FoldChange entspricht dem Logarithmus zur Basis 2 des Quotienten der FPKM Werte.

[18]

Tabelle 3.14: Top 10 signifikant verminderte und erhöhte Mengen kleiner RNA pro *Repeat*-Region

Region	Wert_1	Wert_2	log_2FoldChange	p-value
DIRS_1:2-4824	27172,1	22042,9	-0,301812	0
skipper:0-6994	185,757	1093,41	2,55735	0
DDT_S:0-756	128,979	597,509	2,21182	0
DGLT_P:0-6014	118,477	368,28	1,6362	0
TRE5_B:0-5771	173,582	497,998	1,52052	0
TRE3_D:0-4616	193,285	547,641	1,5025	0
TRE5_A_ModA:0-773	38,616	108,912	1,49589	0
DGLT_A1:2-5048	132,362	368,659	1,4778	0
TRE3_A:2-5248	238,093	657,515	1,4655	0
DDT_B:0-5471	66,298	168,883	1,34899	0
thug_T:0-1130	73,1407	177,902	1,28234	0
TRE5_A1:2-5649	202,063	475,025	1,2332	0
TRE5_C:0-884	202,259	458,852	1,18183	0
Tdd_4:0-3843	83,0176	178,179	1,10183	0
DDT_A:3-5169	67,5791	143,022	1,08159	0
TRE3_B:0-5289	470,062	983,159	1,06457	0
TRE3_C:0-4740	169,308	326,228	0,946233	0
thug_S:0-2192	109,372	189,565	0,793452	0
Tdd_5:0-3809	117,415	192,21	0,711058	0

[18] Die Anzahl aller untersuchten *Reads* innerhalb der *Repeats* sind signifikant reduziert oder erhöht. FPKM Wert_1 entspricht dem Wildtyp und FPKM Wert_2 dem Gendeletionsstamm.

Bei Betrachtung der Anzahl überlappender *Reads* mit annotierten Genen konnten 472 signifikante Unterschiede zwischen Wildtyp und Gendeletionsstamm festgestellt werden (Abbildung 3.37).

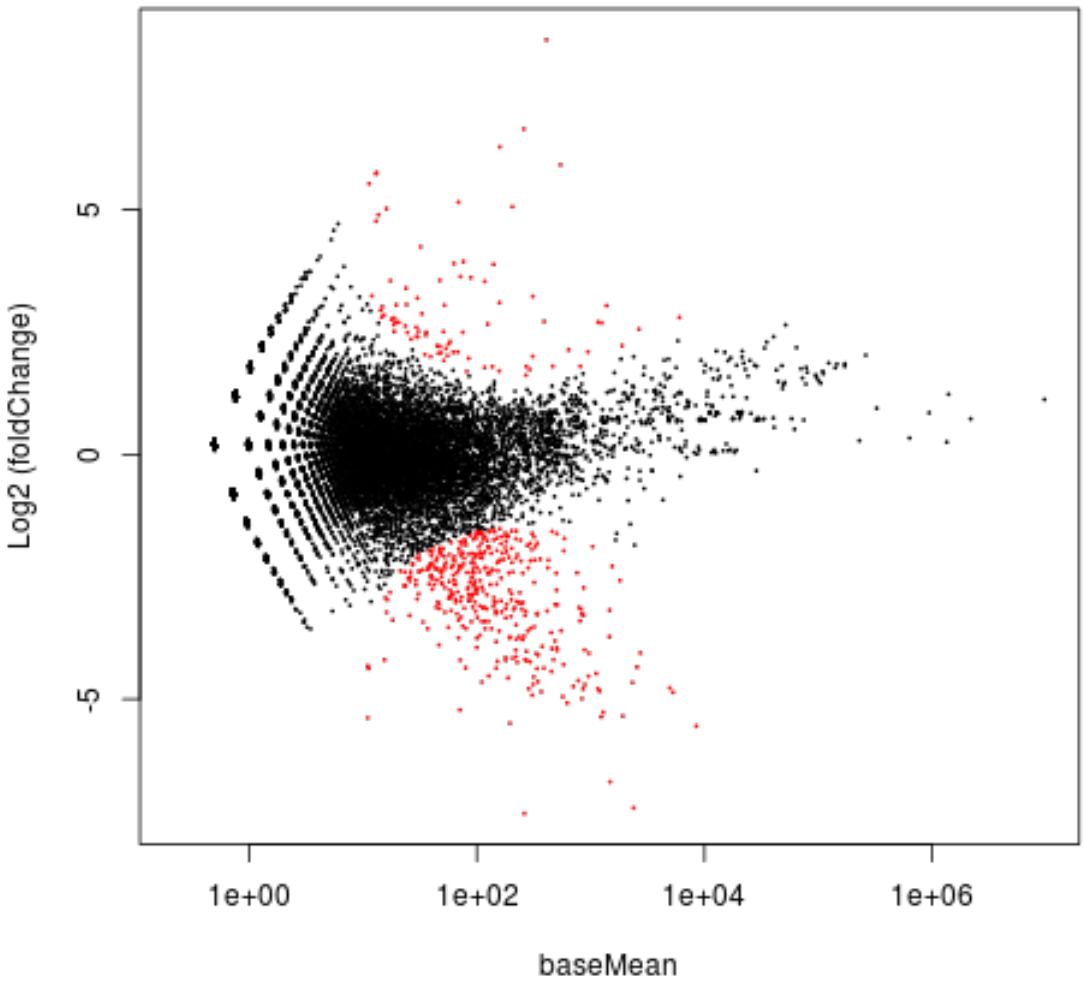


Abbildung 3.37: Signifikante Unterschiede in der Anzahl der *Read*-Überlappungen mit annotierten Genen
Der Signifikanz-Test der Unterschiede der Mittelwerte zwischen Wildtyp und Gendeletionsstamm zeigt 472 siRNA (rot markiert). Die Anzahl der *Read*-Überlappungen mit annotierten Genen ergibt eine 13110 x 4 Matrix. Die jeweiligen Spalten dieser Matrix wurden mit dem Median aus allen Quotienten der Werte und der Spaltenmittelwerte skaliert. Die Werte der X-Achse entsprechen dem Mittelwert (`baseMean`) der Mittelwerte der skalierten Anzahl der *Reads* aus dem Wildtyp (`baseMeanA`) und Gendeletionsstamm (`baseMeanB`). Diese sind gegen die Werte der Y-Achse aufgetragen, die dem Logarithmus zur Basis 2 der Quotienten (`foldChange`) aus `baseMeanB` und `baseMeanA` entsprechen. Ist der P-Value kleiner als $1 * 10^{-5}$ ist der Punkt signifikant und rot markiert.

Tabelle 3.15 zeigt die Top 10 dieser Gene sortiert nach dem P-Value.

[19]

Tabelle 3.15: Top 10 signifikante *Read*-Überlappungen mit Genen

ID	BaseMeanA	BaseMeanB	log_2FoldChange	p-value	AX_1	AX_2	$rrpC_1$	$rrpC_2$
DDB_G0294228_RTE	519,61	3,22	-7,34	1,08e-59	5	1	707	408
DDB_G0294320_RTE	4736,06	31,72	-7,22	2,37e-57	32	28	6415	3750
DDB_G0305714_RTE	2933,11	28,37	-6,69	2,75e-57	22	32	3414	2925
DDB_G0292178_ps	2,29	814,46	8,47	2,73e-54	838	702	2	3
sslA_ps1	5,09	512,98	6,65	4,16e-46	527	443	6	5
DDB_G0294230_RTE	382,05	8,49	-5,49	1,16e-45	10	6	415	413
rpsA	1208,47	35,74	-5,08	4,70e-45	28	40	1467	1140
gpbB	2424,97	58,98	-5,36	5,82e-45	51	61	2705	2545
rpl11	1103,65	35,76	-4,95	8,74e-44	29	39	1351	1029
rpl14	593,00	19,59	-4,92	1,36e-43	21	16	621	666

[19] 3.15 zeigt signifikante mit **DEseq** ermittelte Gene, sortiert nach dem P-Value. Spalten AX_1,AX_2,$rrpC_1$ und $rrpC_2$ zeigen die Anzahl der *Reads*, die diese Gene überlappen.

Abbildung 3.38 zeigt eine *Heatmap* der Top 100 signifikantesten Gene.

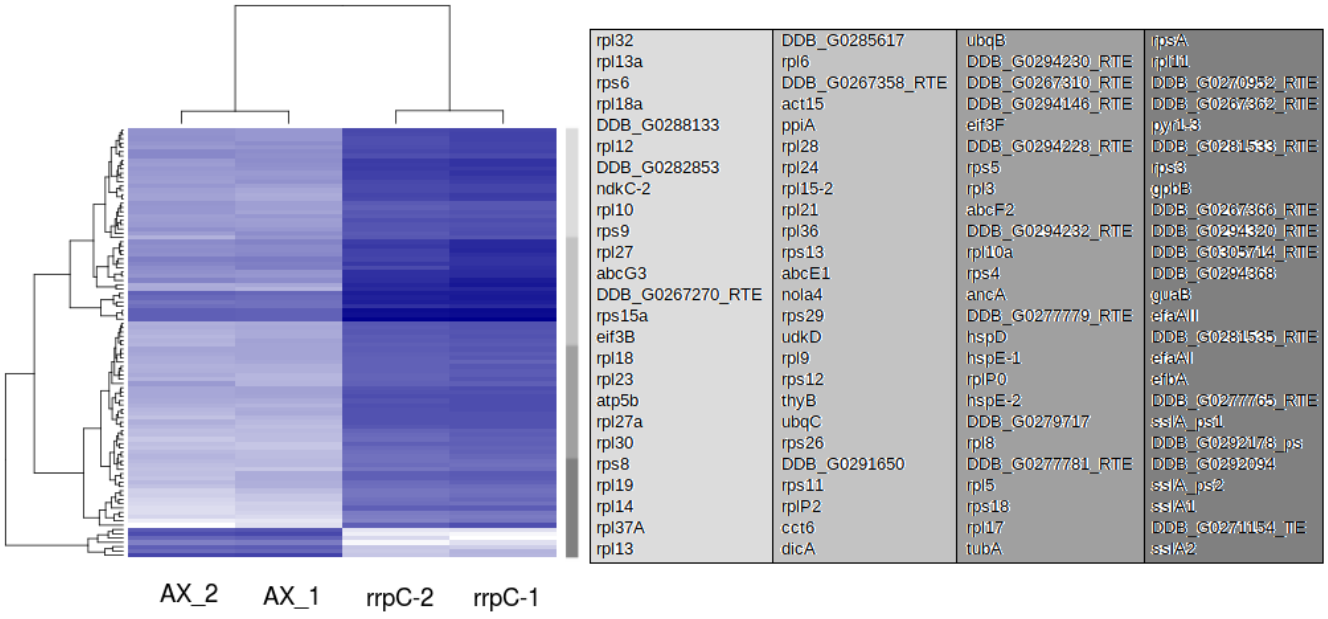


Abbildung 3.38: *Heatmap* der Top 100 signifikantesten Gene

Aus den Top 100, nach dem *P-Value* sortierten Genen, wurde die obige *Heatmap* erstellt. Die Werte der dargestellten Matrix entsprechen der Anzahl überlappender *Reads*, deren Varianz stabilisiert wurde und annähernd konstant ist. Je höher die Anzahl desto blauer ist der Farbton.

4 Diskussion

In diesem Kapitel werden RNAhit, die mit RNAhit durchgeführten Suchen sowie deren Interpretation der Motiventwicklung und das Ergebnis der *Deep Sequencing* Analyse diskutiert.

4.1 RNAhit

4.1.1 Suchprogramme

In RNAhit werden momentan zwei Sequenz- und Strukturprogramme angeboten. Eine Erweiterung ist möglich, wobei die Ausgabe einer Punkt-Klammer-Struktur von Vorteil ist, da diese für die Erzeugung der Faltungseinschränkungen benötigt wird.

Der Test verschiedener Suchprogramme zeigte, dass PatScan [Dsouza et al., 1997] die beste Wahl eines auf *Pattern* basierenden Programmes ist. PatScan ist zwar langsamer als RNAbob, aber dafür werden verschachtelte Motive gefunden. RNAbob findet z. B. das *Pattern* „ACTTGN[6-12]CATA“ in der Sequenz „ACTTGcactggCATAtgCATA“ nur einmal. PatScan ermittelt dagegen zwei Treffer. Trotzdem findet PatScan nicht in jedem Fall alle Treffer, vor allem bei *Mismatches*, Deletionen und Insertionen überlappender Motive [Zanger, 2005]. Falls die Genauigkeit vernachlässigt werden kann, ist speziell für genomweite, große Datenmengen RNAbob zu empfehlen, zumal je nach Filtereinstellung von RNAhit überlappende Treffer im Nachhinein gefiltert werden.

Bis auf RNAfold [Hofacker et al., 1994] waren die nachfolgend getesteten Programme für eine Erweiterung von RNAhit ungeeignet. Aufgrund der strukturellen Konservierung der Motive, jedoch nicht auf Sequenzebene, ist z. B. BLAST als Suchwerkzeug ungeeignet, da variable Helix- und *Loop*-Längen nicht berücksichtigt werden können und die Mindestwortlänge 11 nt beträgt [Altschul et al., 1990]. Ein anderes auf Indizes basierendes Programm aus dem BLAST-Paket heißt seedtop. Dieses Programm ist in der Suche schneller als RNAbob und PatScan, aber die Vor- und Nachbearbeitung der Daten dauert länger. Dazu zählen das Entpacken der Datei, das Vorbereiten der Suche mit formatdb und das Nachbereiten der Treffer in Form der Bestimmung der Sequenz anhand der von seedtop ausgegebenen Positionen und das Erstellen der Punkt-Klammer-Struktur. Des Weiteren können keine komplementären Basenpaare definiert werden und die Suche auf dem Minus Strang geschieht durch einen zweiten Aufruf. Ein anderes Programm, das sich für die Hochdurchsatz-Motivsuche in RNAhit nicht eignet, ist Locomotif [Reeder et al., 2007], da es auf einem Client-Server Prinzip basiert und nicht lokal installiert werden kann. Ein Kriterium für die Aufnahme in RNAhit ist die erfolgreiche Suche nach bekannten Motiven. Da Vmatch z. B. die HHRz in *Arabidopsis thaliana* auf Chromosom 4 [Przybilski et al., 2005] nicht finden konnte, wurde es ebenfalls verworfen. Perreault [Perreault et al., 2011] nutzte für seine Homologiesuchen Infernal [Nawrocki et al., 2009]. Es verwendet das Stockholm-Format, welches zusätzlich neben der Sequenz weitere Eigenschaften, wie die Sekundärstruktur, enthält. Vorteil des Programmes ist, dass es mit dem generierten Kovarianz Modell auch nicht konservierte Sequenzpositionen berücksichtigt und somit natürliche Varianten des Motivs finden kann. Der Nachteil des Programmes ist jedoch, dass es im Vergleich zu anderen sehr langsam und für genomweite Suchen in einem Sequenzraum der Größe $1,55 * 10^{11}$ nt ungeeignet ist.

4.1.2 Faltungsprogramme

In RNAhit werden zwei Faltungsprogramme zur Filterung der primären Treffer angeboten. Die Anwendung thermodynamischer Auswahlkriterien zur Bestimmung der korrekten Faltung wurde bereits von anderen Gruppen demonstriert [Reeder & Giegerich, 2009]. Zur Erweiterung der Auswahl in RNAhit wurden ebenfalls verschiedene Programme getestet. Der Vergleich zeigt Übereinstimmungen und Tendenzen, aber auch Unterschiede. So weichen die ΔG_{free} Berechnungen von kinefold stark von denen durch Mfold berechneten freien Energien ab, was an den verwendeten Energieparametern liegen könnte. Ein wichtiges Kriterium für die Aufnahme in RNAhit ist die Definition eigener *Constraints* und Berechnung der freien Energie eigener erzwungener Strukturen. Da Mfold bereits seit 1995 zur Faltungs- und Energieberechnung in mehreren Publikationen eingesetzt wurde (laut Web of Science http://wokinfo.com/ 3730 Publikationen, Stand 16.04.2012), ist es auch in dieser Arbeit das am häufigsten verwendete Programm.

4.2 Suchergebnisse

In diesem Abschnitt werden die RNAhit-Filter und die durchgeführten Suchen diskutiert. Die gefundenen und gefilterten Suchergebnisse setzen die Annahme voraus, dass gleiche Strukturen die gleiche Funktion bedeuten. Im Fall der Ribozyme ist dies die katalytische Aktivität. Um gleiche Strukturen zu ermitteln, wurden verschiedene Filter eingesetzt. Der Filter der minimalen freien Energie entfernt dabei alle Motive, die eine zu geringe Stabilität besitzen und sich wahrscheinlich nicht in die vorhergesagte Struktur falten. Mögliche redundante Sequenzen aufgrund redundanter Quellen wurden von RNAhit durch einen Überlapp- und *Unique*-Filter entfernt. Der „Haupt"-Filterschritt, bei dem die meisten Sequenztreffer verworfen werden, ist der $\Delta\Delta G$ Wert. Durch diesen ist es möglich anhand des Vergleiches freier Energien zu überprüfen, ob sich eine Sequenz in die gesuchte Struktur falten kann. Außerdem können im Nachhinein Berechnungsungenauigkeiten von Mfold beeinflusst werden, da Mfold bei der Berechnung der erzwungenen freien Energie die tertiären *Loop*-Interaktionen nicht berücksichtigt. Allerdings verdeutlicht Abbildung 3.14 ein Problem der Verwendung der minimalen freien Energie als Filter. Falls eine alternative Struktur existiert, die den gleichen ΔG Wert besitzt wie die erzwungene Struktur, ist dies ein FP Treffer, da Energien verglichen werden und nicht drei dimensionale Koordinaten. Deshalb sollten z. B. zusätzlich die Punkt-Klammer-Strukturen miteinander verglichen werden. Des Weiteren kann in seltenen Fällen ein negativer $\Delta\Delta G$ Wert beobachtet werden. Das bedeutet, dass die freie Energie der erzwungenen Struktur kleiner ist als die der ungezwungenen Struktur. In diesen Fällen wird in der freien Faltung ein mögliches Basenpaar (z. B. HHRz III, Helix III, Position 1 mit der letzten Position) nicht gebildet und besitzt dadurch im Vergleich zur erzwungenen Struktur einen größeren ΔG Wert.

Die Filterreihenfolge spielt ebenfalls eine wichtige Rolle. So ist es entscheidend, dass ΔG_{free} vor $\Delta\Delta G$ gefiltert wird und nicht umgekehrt, da sonst TP entfernt werden. Patscan findet z. B. in *Arabidopsis thaliana* auf Chromosom 4, 3+2 überlappende HHRz III unterschiedlicher Länge an den zwei bekannten HHRz-Positionen [Przybilski et al., 2005]. Wenn zuerst nach dem $\Delta\Delta G$ Wert gefiltert wird und der Treffer mit dem kleinsten $\Delta\Delta G$ Wert bleibt, besitzt die verbleibende Sequenz einen zu kleinen ΔG_{free} Wert und wird im zweiten Filterschritt entfernt. Dadurch geht ein TP verloren. Wird im Gegensatz dazu zuerst nach ΔG_{free} gefiltert, wird diese Sequenz sofort entfernt und es verbleibt ein anderer zweitbester $\Delta\Delta G$ Wert, der nicht gefiltert wird, wodurch die 2 TP erhalten bleiben.

Die automatische Parameterauswahl der Filtereinstellungen zeigte das beste Verhältnis bei 26 TP und 0 FP. Dies ist mit 15% bei 173 möglichen TP eine schlechte Parametereinstellung. Außerdem sind vermutlich unter den 26 TP überlappende Treffer, da der Überlapp-Filter deaktiviert wurde. Der einzige Vorteil ist, dass mit den ermittelten Parametereinstellungen keine FP identifiziert werden. Für eine bessere automatische Parameterwahl werden weitere bekannte TP und vor allem FP benötigt. Die letztendlich ausgewählten Parametereinstellungen für RNAhit wurden mit Hilfe experimenteller Ergebnisse angepasst. Wird der $\Delta\Delta G$ Wert als Filtergrenze zu klein gesetzt, werden bereits als positiv getestete Sequenzen gefiltert.

Als weitere Kontrolle der Filtereinstellungen wurde in 2877 Einträgen (Stand 11.04.2010) der „subviral" Datenbank gesucht [Rocheleau & Pelchat, 2006], die mehr als 100 einzigartige HHRz III Sequenzen aus Satelliten RNA und Viroiden enthält [Tabler & Tsagris, 2004]. Der Grund, warum nicht alle sub-viralen HHRz III gefunden werden konnten (05.08.2010), liegt in der Abweichung einzelner Sequenzen vom konservierten, katalytischen Zentrum, welche sich durch Insertionen (+1 nt) nach CUGANGA oder *Mismatches* an Position 10.1 und 11.1 in Helix II nach CUGANGA und vor GAAA äußern. Die Sequenzen, bei denen das Motiv zutrifft, werden korrekt ausgegeben, was bedeutet, dass die angewandten Filterschritte für die Trennung zwischen TP und FP geeignet sind. Dennoch konnten Sara Völkel, Florentine Stix und Anne Kalweit experimentell zeigen, dass sich in den gefilterten Treffermengen der genomweiten Suchen weitere FP Treffer befinden, was durch weitere Anpassungen der Filtereinstellungen und Deskriptoren behoben werden kann. Werden die Filtereinstellungen jedoch zu strikt gewählt, werden alle primären Treffer gefiltert, wie am Beispiel von *Caenorhabditis elegans* (28.09.2010) zu sehen. Um die Filtereinstellungen weiter anpassen zu können, müssten ebenfalls die gefilterten Mengen nach TP durchsucht werden, besonders Sequenzen, deren Werte nahe der Filtergrenze liegen. Es könnte aber auch bedeuten, dass in *Caenorhabditis elegans* keine korrekt gefalteten HHRz III mit dieser Motivbeschreibung existieren.

Allgemein gibt es im Vergleich der Suchergebnisse nach HHRz zwischen den einzelnen Forschungsgruppen ([de la Peña & Garcia-Robles, 2010b, Seehafer et al., 2011, Perreault et al., 2011]) signifikante Überschneidungen, da alle das konservierte, katalytische Zentrum des HHRz berücksichtigen. De la Peña und Garcia-Robles fanden HHRz Typ I und III in 178 Eukaryonten und 3 Bakterien [de la Peña & Garcia-Robles, 2010b].

Perreault suchte nach HHRz in allen drei Typen und identifizierte ungefiltert

- 25 Treffer in 14 Viren,
- 2 Treffer in *Cenarchaeum symbiosum* einem Archaeon,
- 160 Treffer in 58 Bakterien,
- und 31101 Treffer in 79 Eukaryonten.

Die Treffer, die sich unterscheiden, stammen entweder aus unterschiedlichen Quellen oder sind aufgrund der verwendeten Methoden und Deskriptoren verschieden.

Je genauer ein Deskriptor ist, desto größer ist die Chance TP zu finden, jedoch mit der Gefahr Variationen des Motivs zu verfehlen. Wird der Deskriptor allgemeiner definiert, steigt die Anzahl der FP. Dies zeigt das Problem der korrekten Deskriptorwahl.

Die mit den Deskriptoren gefundenen Motive werden zur Zeit nach möglichen biologischen Funktionen untersucht. Was alle Ribozyme gemeinsam haben, ist die Teilung einer transkribierten RNA durch das Motiv. Auf diese Weise können zwei miteinander verbundene Sequenzen voneinander getrennt bzw. miteinander verbunden werden. Die Produkte der Spaltreaktion könnten Substrate für darauffolgende Reaktionen sein, z. B. eine anschließende Ligation, Spaltung in *trans* oder eine Protein katalysierte Reaktion [Hammann et al., 2012]. De la Peña und Garcia-Robles konnten z. B. die Fusion eines durch ein HHRz gespaltenen Introns mit einer *small nuclear RNA* (snRNA) beobachten [de la Peña & Garcia-Robles, 2010a]. Der nach der Spaltung entstehende Überhang könnte der Integration in die DNA dienen, vorausgesetzt diese besitzt einen komplementären Überhang. Diese und weitere mögliche Funktionen werden in dem *Review* [Hammann et al., 2012] zusammengefasst. So könnten die Ribozyme durch die Spaltung die Produktion von *Piwi interacting RNA* (piRNA) oder anderen kleinen RNA hemmen oder die Generierung regulatorischer RNA unterstützen. Ribozyme könnten auch der polycystronischen Prozessierung von *transfer RNA* (tRNA) dienen, ähnlich wie in marinen Metagenomen. In Bakterien könnte ein durch ein Ribozym erzeugter *Frame Shift* ein neues Leseraster erzeugen und somit eine bakterielle Form des alternativen *Splicings* darstellen. Eine weitere mögliche Funktion könnte die Inhibition des RNAi-Mechanismus sein. Dazu würde, wie bei dem Pilz *Yarrowia lipolytica*, die komplementäre RNA den Doppelstrang nicht bilden, da diese zuvor gespalten wird [Hammann et al., 2012].

In den folgenden Abschnitten wird detailliert auf die Suchergebnisse der einzelnen Ribozyme eingegangen und deren Bedeutung diskutiert.

4.2.1 Hairpin Ribozyme

Die Suchen nach *Hairpin* Ribozymen ergaben unabhängig vom Suchprogramm und mit bzw. ohne Wobble Basenpaare eine *Hairpin* Ribozym ähnliche Sequenz in *Monodelphis domestica*, bei der jedoch weitere Teile des Motivs fehlen. Dies ist ebenfalls bei der Sequenz aus *Pongo pygmaeus* der Fall. Bei beiden Sequenzen handelt es sich wahrscheinlich um Zufallstreffer, da der Deskriptor eine Wahrscheinlichkeit von rund $P_{Motiv} = \left(\frac{6}{16}\right)^5 * 1 * \left(\frac{1}{4}\right)^4 * \left(\frac{6}{16}\right)^4 * \left(\frac{1}{4}\right)^2 * 1 * \left(\frac{1}{4}\right)^8 * 1 = 5.46313 * 10^{-13}$ besitzt. Es sind demnach 523 bzw. 589 Treffer zufällig in den Genomen *Monodelphis domestica* und *Pongo pygmaeus* zu erwarten.

In einer weiteren Suche konnten die aus subviral DB bekannten *Hairpin* Ribozyme mit dem Deskriptor aus Abbildung 2.11(d) gefunden werden. Diese wurden jedoch durch den $\Delta G_{ratio} > 0.6$ Filter entfernt, was bedeutet, dass der Filter zu hoch gewählt wurde und dadurch mögliche weitere TP der primären Treffermenge gefiltert wurden.

Die im Vergleich zum HHRz gefundene Menge ist sehr gering, was auf eine unabhängige Verbreitung deutet, obwohl z. B. in der *Rolling circle* Replikation sub-viraler Pathogene beide Ribozyme vorkommen. Die Suche nach *Hairpin* Ribozymen auf dem komplementären Strang der 160 neuen HHRz [Seehafer et al., 2011] ergab keine Treffer. Demnach scheint die Kombination aus *Hairpin*- und *Hammerhead* Ribozym spezifisch für die *Rolling circle* Replikation zu sein.

Das Vorkommen in sub-viralen Pathogenen deutet auf einen frühen Ursprung dieser Ribozyme.

4.2.2 Hammerhead Ribozyme

Unter allen gesuchten Ribozymen kommen *Hammerhead* Ribozyme am häufigsten vor. Dies zeigte bereits die erste Suche nach Typ III (10.07.2009) in einer ersten Anwendung der *Pipeline* und ergab Treffer in mehreren Bakterien und Eukaryonten. Spätere genomweite Suchen zeigten, dass HHRz-Sequenzen in sub-viralen Pathogenen und in allen Bereichen des Lebens gefunden werden, wobei sie in Archaeen unterrepräsentiert sind. Die unterschiedlichen HHRz-Typen kommen dabei zum Teil in ähnlichen Mengen vor. Dies deutete eine genomweite Suche an mit 549262 HHRz II Treffern (19.3.2010) und 580196 HHRz III Treffern (10.11.2009) aus allen Bereichen des Lebens und wurde von Perreault für Bakterien bestätigt [Perreault et al., 2011]. Des Weiteren konnte in repetitiven Regionen aus *Xenopus tropicalis* ein modularer Aufbau beobachtet werden, in dem alle drei Typen innerhalb einer Region möglich sind. Es ist vorstellbar, dass je nach benötigter Schnittstelle der jeweilige Typ aktiv ist. Daraus ergeben sich $\left(2^3 = 8\right)$ 7 Möglichkeiten den Sequenzabschnitt zu unterteilen, vorausgesetzt der gesamte Abschnitt wird transkribiert und liegt als RNA vor. Eventuell wird nur ein Teilabschnitt transkribiert und zwei oder eine Spaltstelle verwendet. Am häufigsten wurden HHRz als Typ I gefunden. Die Mehrheit der Treffer in Metazoa sind vom Typ I und Typ III entspricht in diesen eher einer Ausnahme [Hammann et al., 2012].

Die Auswertung der *Mutual Information* (MI) zeigte, dass Typ I HHRz eine größere Übereinstimmung co-evolvierender Positionen mit sub-viralen Sequenzen besitzen als Typ III HHRz [Hoffgaard et al., Prep]. Das heißt, bei einem vorausgesetzten gemeinsamen Ursprung, dass Typ I im Vergleich zu Typ III älter sein könnte, was ebenfalls durch die hohe Häufigkeit von Typ I Treffern bekräftigt wird. Die unterschiedlichen Treffermengen der einzelnen Typen könnte auch bedeuten, dass sie einem unterschiedlichen selektiven Druck unterliegen, bei gleichem Alter. Die signifikanten co-evolutionären Zusammenhänge bestehen je nach Typ zwischen den Nukleotidpositionen der Helices I und III. Dies spricht trotz variabler Sequenz für eine Konservierung auf Strukturebene, die bei Änderung einer Helixposition die Änderung der anderen bedingt, obwohl keine physische Interaktion besteht. Ein Großteil der HHRz I Kandidaten besitzt eine Länge zwischen 50 bis 60 nt, was der ursprünglichen Länge der HHRz I entsprechen könnte, vorausgesetzt sie besitzen einen gemeinsamen Ursprung. Größere Längen könnten durch Insertionen in *Loop*- und Helixregionen unter Beibehaltung der Struktur entstanden sein. Die charakteristischen *Loop*-Größen in HHRz I von 5, 7 und 9 nt deuten auf Interaktionen über Kissing Komplexe.

Da unterschiedliche Variationen des katalytischen Zentrums bekannt sind [Perreault et al., 2011] wurde der Deskriptor modifiziert und z. B. an den Positionen 3 und 8 verallgemeinert (03.09.2009, 26.03.2010). Das Ergebnis der Verallgemeinerung zeigte ein fast dreifach erhöhtes Vorkommen von U3A8 im Vergleich zu C3G8, obwohl *in vitro* gezeigt wurde, dass diese Variation eine verminderte selbst-spaltende Aktivität besitzt [Przybilski & Hammann, 2007]. Dennoch scheint dies die am häufigsten verwendete HHRz III Variante zu sein, was evolutionsbiologisch betrachtet innerhalb der Typ III HHRz auf ein höheres Alter deutet, falls sie einen gemeinsamen Ursprung besitzen. Diese und andere Variationen der HHRz sprechen für eine Entwicklung des Motivs und Optimierung der Spaltungsrate in der Evolution. Allgemein könnten beobachtete Variationen (zusammengefasst in [Hammann et al., 2012]) eine Anpassung an Umweltbedingungen sein [Perreault et al., 2011]. Eine mögliche neue Variation des katalytischen Zentrums wurde in *Arabidopsis lyrata* entdeckt. Eine BLAST Ähnlichkeitssuche des neu gefundenen HHRz III Arth1 aus *Arabidopsis thaliana* identifizierte die Sequenz. Zu den Ribozymen „Ara1“ und „Ara2“ auf Chromosom 4 von *Arabidopsis thaliana* [Przybilski et al., 2005] existiert in der nah verwandten Art *Arabidopsis lyrata* ein Homolog auf Chromosom 7. Die *Synteny* (Abbildung 3.17) zeigt, dass ein großer Teil auf Chromosom 7 abgebildet werden kann. Die Gene im Umfeld der homologen Sequenz in *Arabidopsis lyrata* sind unbekannt. *XM_002869310.1* kodiert für ein hypothetisches Protein (93 Aminosäuren). Ein lokales Alignment mit den Proteinen AT4G30860.1 (497 Aminosäuren) und AT4G30870.1 (659 Aminosäuren) aus *Arabidopsis thaliana* ergab auf einer Länge von 13 und 64 Aminosäuren eine Ähnlichkeit von 53.8% und 40.6%. Das heißt, dass die Proteine aus *Arabidopsis thaliana* zu dem hypothetischen Protein relativ verschieden sind, was wiederum bedeutet, dass das unmittelbare genetische Umfeld der homologen Sequenz verschieden ist. Die gefundene Variation zeigt, dass weitere alternative Strukturen, die mit bisherigen Deskriptoren nicht gefunden wurden, möglich sind. Ein weiteres Beispiel ist die von Anne Kalweit als katalytisch aktiv gezeigte Variation des katalytischen Zentrums UG, CUGANGA, GAAA, welche in 3 Eukaryonten gefunden wurde.

In einer anderen Suche mit der N3N8 Verallgemeinerung wurden neben den Watson-Crick Basenpaaren Wobble Basenpaare zugelassen (16.09.2009), was die primäre Treffermenge in den Schistosomen mehr als verdreißigfachte. Dies könnte an der Wahrscheinlichkeit liegen, die theoretisch $\left(\frac{6}{16}\right)^{12} / \left(\frac{4}{16}\right)^{12} = 129.7463$ rund 130 mal mehr Treffer erwarten würde. Die zweite genomweite Suche nach HHRz III erfolgte mit dem verallgemeinerten Deskriptor am 10.11.2009, bei der erstmals eigene Faltungsbeschränkungen definiert wurden. In der dritten genomweiten Suche vom 26.03.2010 kam zu dem verallgemeinerten Deskriptor und den Faltungsbeschränkungen ein weiterer Filter hinzu (ΔG_{free}). Dieser bestätigt in der Suche vom 16.07.2010 die zwei bekannten HHRz III aus *Arabidopsis thaliana* [Przybilski et al., 2005] und einen dritten Treffer, der bereits durch Anne Kalweit als katalytisch aktiv charakterisiert wurde.

In einer vierten genomweiten Suche (02.09.2010) wurde nach der minimalen HHRz III Struktur gesucht, ohne alternative Helixlängen, mit einem jeweils auf 200nt erweiterten *Loop* und sehr strikten Parametereinstellungen ($\Delta\Delta G = 0$). Die *Loop*-Größen wurden anschließend jeweils auf 1000nt erweitert (12.11. und 03.12.2010). Es war zu erwarten, dass aufgrund der Substratlänge in *trans* spaltende HHRz mit erweitertem *Loop* I häufiger vorkommen. Dies wurde bestätigt, da bei gleichem Suchraum und Parametereinstellungen mehr Treffer gefunden wurden als bei der HHRz III *Loop* II Erweiterung. Die Wahrscheinlichkeit beider Erweiterungen in Zufallssequenzen ist annähernd gleich. Große *Loops* könnten z. B. mögliche Proteinbindestellen enthalten [Hammann et al., 2012].

Die besonders hervorgehobenen 160 neuen HHRz III der Veröffentlichung [Seehafer et al., 2011] wurden näher untersucht. In den eukaryontischen Genomen, in denen die meisten Treffer gefunden wurden, wahrscheinlich auch zufällig aufgrund der Größe des Suchraumes, variiert die Anzahl und die Lokalisierung der Treffer sehr stark. Zum Teil wurde nur ein einzigartiges HHRz in einem Organismus gefunden, wie z. B. in *Aedes aegypti* auf supercont1.304 Position 671679 bis 671786 *Assembly* `AaegL1`. Die Einzeltreffer deuten, je nach angenommener Motivwahrscheinlichkeit und Genomgröße, auf Zufallstreffer hin. Dies bedeutet nicht zwangsläufig, dass sie katalytisch inaktiv sind, da in *Systematic Evolution of Ligands by Exponential Enrichment* (SELEX) Experimenten unter annähernd physiologischen Bedingungen gezeigt werden konnte, dass aus einer Zufallssequenz heraus ein katalytisch aktives *Hammerhead* Ribozym entstehen kann [Salehi-Ashtiani & Szostak, 2001]. Dies könnte in der Evolution zu einem selektiven Vorteil geführt und dadurch das Motiv weiter verbreitet haben. Darum besteht die Annahme [Uhlenbeck, 1987, Perreault et al., 2011, Jimenez et al., 2011], dass HHRz unabhängig voneinander mehrmals entstanden sein könnten, da sie außerdem weit verteilt sind, in verschiedenen Sequenzregionen vorkommen, sehr variabel sind und große Längenunterschiede aufweisen. Teilweise wurde ein Treffer mehrfach gefunden, weil dieser auf unterschiedlichen *Scaffolds* liegt, wie z. B. beim Elefanten *Loxodonta africana* auf *Scaffold* 160, 162, 188 und 189. Dies hat jedoch keinen biologischen Grund, sondern liegt an dem unvollständigen *Assembly*, da eine Position durch mehrere *Contigs / Scaffolds* überlappt und ein Treffer dadurch mehrmals gefunden wird. Da es in anderen Organismen wiederum Hunderte und Tausende einzigartige HHRz-Treffer gibt (z. B. HHRz I in *Schistosoma mansoni*), könnte dies bedeuten, dass in Organismen mit Einzeltreffern weitere alternative Variationen oder Ribozyme existieren. Eine hohe Trefferanzahl muss wiederum nicht heißen, dass alle Treffer katalytisch aktiv sind. Sara Völkel konnte in ihrer Diplomarbeit zeigen, dass sich unter den 10 von ihr getesteten Kandidaten aus den 35 *Hydra magnipapillata* Sequenzen FP Treffer befinden. Dies ist auch bei anderen veröffentlichten, experimentell ungetesteten, großen Treffermengen anzunehmen.

Von einigen der 160 Sequenzen konnten Spaltprodukte in den EST Daten gefunden werden, was für eine *in vivo* Aktivität dieser Treffer spricht. Bis auf wenige Ausnahmen befinden sich die Treffer in repetitiven Sequenzbereichen, in Introns Protein-kodierender Gene oder in intergenischen Regionen. Eine Beobachtung, die De la Peña [de la Peña & Garcia-Robles, 2010b] und Perreault [Perreault et al., 2011] gemacht haben, war, dass HHRz aus intergenischen Bereichen in Bakterien mit Bakteriophagen DNA im Zusammenhang stehen. HHRz könnten auf diese Weise die Integration der Phagen DNA ermöglicht haben. Die Fundorte in repetitiven Regionen stehen zum Teil im Zusammenhang mit Retrotransposons. Die hohe Trefferanzahl in *Schistosoma mansoni* könnte damit erklärt werden, dass das Genom aus 40% repetitiven Sequenzen und über 70 Transposon Familien besteht [Berriman et al., 2009]. In repetitiven Regionen könnten die Ribozyme der Prozessierung von SINEs dienen und somit einheitlich lange Fragmente (Monomere) erzeugen [Hammann et al., 2012], ähnlich wie bei der *Rolling circle* Replikation [Branch & Robertson, 1984]. Die Sequenzähnlichkeiten der HHRz zwischen den Organismen könnten sich also durch mobile Elemente, wie Retrotransposons oder sub-virale Pathogene verbreitet haben [Roychowdhury-Saha et al., 2011]. Die Anwesenheit in Introns könnte ebenfalls der exakten Prozessierung dienen, z. B. von *trans* agierenden regulatorischen RNA Molekülen, ähnlich wie bei der Prozessierung von miRNA aus Introns [Kim & Kim, 2007]. Oder sie könnten für die Kontrolle des alternativen *Splicings* von Bedeutung sein, ähnlich wie bei den intronischen HHRz beschrieben von De la Peña [de la Peña & Garcia-Robles, 2010a].

Das Erstellen eines MSA der 160 HHRz war aufgrund der variablen Sequenzlängen schwierig. Es wurden mit Hilfe Distanz basierter Ansätze zwei phylogenetische Bäume erstellt. Diese zeigen innerhalb von Untergruppen einen zum Teil hohen Grad phylogenetischer Konservierung. Es gibt Cluster aus Arten (z. B. *Hydra magnipapillata*, *Schistosoma mansoni* oder *Tarsius syrichta*), Gattungen (*Arabidopsis*, *Drosophila* und *Aspergillus*) und Klassen (z. B. *Gallus gallus* und *Taeniopygia guttata*) in denen die HHRz ähnlich sind. So können z. B. Treffer in *Arabidopsis thaliana*, *Arabidopsis lyrata* und *Vitis vinifera* in einem Cluster der Gattung Pflanzen gruppiert werden, was auf ein gemeinsames HHRz deutet. Die Gruppierung des Bakteriums *Azorhizobium caulinodans* zusammen mit den Treffern in *Arabidopsis thaliana*, *Arabidopsis lyrata* und *Vitis vinifera* weist auf mobile Elemente zwischen Bakterien und Eukaryonten. *Azorhizobium caulinodans* ist ein Pflanzensymbiont und kommt in der Gattung Sesbania vor [Lee et al., 2008]. Es liegt daher die Vermutung nahe, dass sich in diesen bisher nichtsequenzierten Pflanzengenomen ebenfalls HHRz befinden, ähnlich zu den gefundenen. In einer Ähnlichkeitssuche des Treffers, der in Vögeln gefunden wurde, konnte z. B. in dem neu sequenzierten Genom *Meleagris gallopavo* ebenfalls ein Treffer gefunden werden. De la Peña und Garcia-Robles konnten in Reptilien, Vögeln und Säugetieren auch hochkonservierte HHRz finden, die alle auf dem *Sense* Strang von langen Introns dreier unterschiedlicher Gene liegen [de la Peña & Garcia-Robles, 2010b].

Vorausgesetzt HHRz besitzen einen gemeinsamen Ursprung, dann wäre, den Bootstrap-Werten zu Folge, das Alter der HHRz-Treffer zwischen dem Reismehlkäfer *Tribolium castaneum* und dem Krallenfrosch *Xenopus tropicalis* kleiner als zwischen den Pflanzen *Arabidopsis thaliana* und *Vitis vinifera*. Gibt es keinen gemeinsamen Ursprung könnte es sich um zwei verschiedene HHRz III handeln, die in den genannten Organismen ähnlich sind. In der Heatmap (Abbildung 3.21) besteht der größte Abstand zwischen Myva1 und Vipa8. Eine Ursache ist der große Längenunterschied von 192 nt und eine andere die Anzahl der *Mismatches* an den variablen Positionen. Bei diesen zwei Beispielen könnte es sich ebenfalls um zwei verschiedene HHRz handeln oder um ein HHRz, welches durch Insertions- und Substitutionsereignisse in das andere umgewandelt wurde.

Ein Zusammenhang zwischen Motivlänge und der Genomgröße ergab für HHRz I Treffer keine Korrelation, was bedeutet, dass die Länge des HHRz unabhängig von der Herkunft ist, ob Bakterium oder Eukaryont.

Wie in Tabelle 3.3 und 3.7 zu sehen, existieren je nach Filtereinstellung 8 mögliche HHRz in 4 Archaeen (*Archaeoglobus fulgidus*, *Nitrosopumilus maritimus*, *Sulfolobus islandicus* und *Sulfolobus tokodaii*). Diese sind im Vergleich zu den Bakterien und Eukaryonten weit unterrepräsentiert. Fraglich ist auch, ob die ermittelten Sequenzen eine selbst-spaltende Aktivität besitzen. Zum Einen haben sie hohe $\Delta\Delta G$ Werte größer 1.40 kcal/mol und zum Anderen besteht *Stem* III bei den 5 HHRz I aus einem oder zwei Basenpaaren. Diese enthalten jeweils nur zwei Wasserstoffbrückenbindungen (AU,UA,GU,UG) und würden unter den extremen Umgebungstemperaturen von 80°C und höher (*Archaeoglobus fulgidus* [Klenk et al., 1997]) nicht in die vorhergesagte Struktur falten [Vazquez-Tello et al., 2002]. Selbst-spaltende HHRz konnten bis zu einer Temperatur von 60°C beobachtet werden [El-Murr et al., 2012]. Bei höheren Temperaturen überwiegt die Hydrolyse [El-Murr et al., 2012]. Falls die Sequenzen FP sind, existiert eventuell zwischen Bakterien und Eukaryonten ein gemeinsamer Vorfahr mit einem ersten HHRz dessen Struktur in der Evolution vererbt wurde. Dann ist es jedoch unwahrscheinlich, dass die HHRz aus einer RNA-Welt stammen. Die Abwesenheit von HHRz könnte ein Indiz für mögliche alternative Variationen, Ribozyme oder andere von der Zelle genutzte Strukturen und Mechanismen sein oder das langsame Verschwinden des Motivs bedeuten, da die Funktion z. B. von Proteinen übernommen wird, wie z. B. durch Ribonuklease A [Kartha, 1967]. Ein anderer relativ unwahrscheinlicher Grund für die Abwesenheit der HHRz könnten, angesichts heutiger präziser Techniken (siehe Abschnitt 2.4.18), Sequenzierungsfehler sein, so dass die vorhandenen Sequenzen fehlerhaft sind und deshalb die Motive nicht gefunden werden.

4.2.3 Hepatitis Delta Virus Ribozyme

Die leere Treffermenge mit dem Deskriptor aus Abbildung 2.12(a) [Webb et al., 2009] kann bedeuten, dass es in *Dictyostelium discoideum* keine Hepatitis Delta Virus (HDV) ähnlichen Ribozyme gibt oder dass sie in bisher unbekannter Form existieren. Um Variationen zu finden, wurde der Deskriptor verallgemeinert. Die Gefahr besteht, dass dadurch FP Treffer ermittelt werden. Ob es sich bei den gefunden Sequenzen um neue HDV ähnliche Ribozyme handelt, kann letztendlich nur experimentell mit Gewissheit gesagt werden. Die Suchen wurden manuell ohne RNAhit durchgeführt, da zur Zeit nur UNAfold und Mfold implementiert sind, die keine Pseudoknoten vorhersagen und somit die erzwungene HDV-Struktur nicht berechnen können. Bei den Treffern, die mit dem Deskriptor aus Abbildung 2.12(d) gefunden wurden, handelt es sich bei *ublcp1-1* um ein Gen, das für Ubiquitin-Like domain-containing CTD Phosphatase 1 kodiert und eine zweite Kopie *ublcp1-2* besitzt sowie um ein WD40 *Repeat* und ein Pseudogen für eine beta-ketoacyl Synthase. Ein globales *Alignment* des DIRS-1 Treffers gegen die DIRS-1 Konsensussequenz zeigte eine Zuordnung zum 3' Ende, ähnlich wie bei Ruminski *et al.*, die HDV ähnliche Ribozyme in zahlreichen LINEs finden konnten [Ruminski et al., 2011]. Repetitive Regionen könnten daher ein weiteres Filterkriterium sein.

Die Struktur dieser Ribozyme ist komplexer als die der HHRz. Sie kommen ebenfalls weit verbreitet in der Natur vor, konnten jedoch in SELEX Experimenten nicht nachgewiesen werden [Webb et al., 2009, Webb & Lupták, 2011, Ruminski et al., 2011]. Das heißt eine zufällige Entstehung ist unwahrscheinlich und deutet auf eine Entwicklung aus einem einfacheren Motiv, wie dem HHRz. Es fehlt jedoch die Ähnlichkeit zu anderen bekannten Motiven, was für weitere bisher unbekannte Ribozyme spricht.

4.2.4 Varkud Satellite Ribozym

Eines der komplexesten Ribozyme, die bisher entdeckt wurden, ist das *Varkud Satellite* Ribozym. Bisher ist nur eine einzige Sequenz bekannt. Eine Möglichkeit ist, dass sich das Motiv ebenfalls aus einem einfacheren Ribozym entwickelt hat. So ähnelt der katalytische Mechanismus dem der *Hairpin* Ribozyme [Wilson & Lilley, 2011]. Die geringe Häufigkeit komplexer Ribozyme könnte ein Anzeichen für die Verdrängung und Ersetzung durch einfachere Motive wie dem HHRz sein. In *Neurospora crassa* konnten 113 primäre HHRz I (07.01.2011) und 91 primäre HHRz III gefunden werden (10.11.2009).

4.3 Multiple Sequence Alignment

Das Alignieren von RNA-Sequenzen ist ein schwieriges Problem, da die Ähnlichkeit zum Teil ausschließlich auf der Struktur beruht. Bestes Beispiel sind die 160 HHRz III Sequenzen. Durch die große Heterogenität in den Sequenzen und der Sequenzlänge ist es mit den getesteten *Alignment*-Algorithmen nicht gelungen ein MSA zu erzeugen, in dem das konservierte, katalytische Zentrum korrekt untereinander aligniert wurde. Durch das Entfernen der *Loop*-Region und blockweise Alignieren der Teilsequenzen konnte ein besseres MSA erstellt werden.

4.4 Motivwahrscheinlichkeit

Für die Berechnung der Wahrscheinlichkeit, dass ein gegebenes Motiv in einer Sequenz vorkommt, gibt es mehrere Ansätze [Kennedy et al., 2008]. Gelingt es diese abzuschätzen, können Aussagen über die Evolution des Motivs getroffen werden. Durch die Variabilität der Motive speziell der HHRz in Sequenz und Struktur wurde die Wahrscheinlichkeit mit Hilfe von Zufallssequenzen und anschließender Ermittlung der Trefferanzahl auf $3.63 * 10^{-8}$ geschätzt. Das Auftreten in kleineren Genomen als 110 Millionen Basenpaare (MBp) und die 10 mal höhere gefundene Trefferanzahl ist somit kein Zufall.

4.5 Deep Sequencing

In diesem Abschnitt werden die Ergebnisse des zweiten Projektes diskutiert. Es wurden kleine RNA eines Wildtyps und eines *rrpC* Gendeletionsstamms von *Dictyostelium discoideum* sequenziert und ausgewertet. Zelluläre RdRP werden z. B. im RNAi-Mechanismus zum *Silencing* von Genen verwendet und sind an verschiedenen Prozessen beteiligt. Sie produzieren komplementäre Sequenzen für eine Ziel RNA über einen Dicer-abhängigen und einen Dicer-unabhängigen Mechanismus, welche daraufhin abgebaut wird [Maida & Masutomi, 2011]. Aus früheren Arbeiten ist bekannt, dass im *rrpC* Gendeletionsstamm auf Transkript-Ebene die Anzahl der Retrotransposon mRNA von DIRS-1 stark erhöht ist [Kuhlmann et al., 2005]. Für das Retrotransposon Skipper konnte im *rrpC* Gendeletionsstamm kein signifikanter Anstieg der mRNA festgestellt werden [Wiegand et al., Prep]. Aus Northern Blot Experimenten (von Stephan Wiegand) ist außerdem bekannt, dass im *rrpC* Gendeletionsstamm die Menge kleiner RNA, vor allem für DIRS-1, stark reduziert ist [Wiegand et al., Prep]. Dies wurde durch die *Deep Sequencing* Daten bestätigt, die eine Reduktion kleiner RNA im *rrpC* Gendeletionsstamm um 42% zeigen. Die Reduktion wurde ebenfalls in anderen früheren Arbeiten beobachtet [Hinas et al., 2007]. Das heißt, durch die fehlende RrpC wird die dsRNA nicht gebildet und somit die siRNA nicht prozessiert, welche sonst an der Spaltung der mRNA des Transposons beteiligt sind. Damit kommt es zu einem Anstieg der mRNA im Deletionsstamm. Was wiederum bedeutet, dass die Anreicherung nicht das Ergebnis einer erhöhten Transkription ist, sondern eines reduzierten Abbaus. Der Unterschied zwischen Wildtyp und *rrpC* Gendeletionsstamm könnte zudem einen technischen Hintergrund haben, so dass z. B. unterschiedliche RNA-Mengen in der Sequenzierung geladen bzw. amplifiziert wurden. Dies kann jedoch, aufgrund der zweifachen Bestätigung in den Replikaten, vernachlässigt werden.

Zu Beginn der Sequenzierung wurde an das 3' Ende der RNA eine Adaptersequenz ligiert, um die Sequenz auf einer „*Flow Cell*“ Oberfläche zu binden. Teile dieser Adaptersequenz können in den *Reads* enthalten sein. Nach deren Entfernung besitzt ein Großteil der gekürzten *Reads* eine Länge von 21 nt. Hinas *et al.* identifizierten ebenfalls eine große Menge 21 nt langer RNA, die überwiegend aus DIRS-1 Retrotransposons stammen [Hinas et al., 2007]. In den Sequenzierungsdaten sind es im Wildtyp durchschnittlich 33.2% aller *Reads* und im Gendeletionsstamm 19.5%, die DIRS-1 zugeordnet werden können und 21 nt lang sind. Dabei handelt es sich wahrscheinlich um siRNA oder miRNA.

Bei dem auffälligen Peak in Abbildung 3.30(d) bei einer Länge von 10 nt im $rrpC_2$, könnte es sich um ein Spaltprodukt handeln. Da der Peak jedoch im $rrpC_1$ nicht erhöht ist, scheint es ein spezifisches Merkmal des Replikats zu sein.

Die hohe Anzahl an Duplikationen in den Replikaten liegt vermutlich daran, dass viele *Reads* aus repetitiven Regionen stammen und daher wenig Variabilität zeigen.

Die Gefahr beim *Mapping* mit *Alignment*-Raten über 90% besteht darin, dass viele FP *Reads* abgebildet werden, die ihren Ursprung an anderen Positionen besitzen. Aus diesem Grund wurden die Standardeinstellungen für das *Mapping* gegen das *Dictyostelium discoideum* Genom und -M 500 -N 1 -L 20 -R 3 -D 20 -i S,1,0.50 für das *Mapping* gegen die *Repeat*-Bibliothek ausgewählt. Im AX2 Wildtyp können kleine RNA überwiegend ribosomalen RNA Genen, Protein-kodierenden Genen und repetitiven Sequenzen, einschließlich mobilen Elementen zugeordnet werden. Die meisten *Reads* des Wildtyps wurden gegen Chromosom 1 aligniert. Auf Chromosom 1 gibt es 1941 annotierte Gene. Davon kodieren 39 für DIRS-1 Fragmente und 10 für Skipper. Im Gendeletionsstamm wurden die meisten *Reads* gegen das extra-chromosomale rDNA-Palindrom aligniert, auf dem bisher nur Gene für rRNA bekannt sind. Die Reduktion der *Read*-Überlappung mit repetitiven Sequenzen und allgemein die Verteilung kleiner RNA im Vergleich zum Wildtyp zeigt Abbildung 3.33. Das *Mapping* zeigt zwei spezielle Regionen, denen die *Reads* auf dem extra-chromosomale rDNA-Palindrom zugeordnet werden konnten (Abbildung 3.36). Zum einen an den Positionen der annotierten rRNA Gene und zum anderen auf dem *Sense* Strang im 3' Bereich, der eine hohe Ähnlichkeit (99%) zu den 5' Positionen besitzt und eine Kopie sein könnte, die alternativ genutzt wird [Boesler et al., 2011]. Hinas *et al*. berichteten von einer auffälligen Region im intergenischen Bereich zwischen dem 26S rRNA und dem 5S rRNA Gen [Hinas et al., 2007]. In Abbildung 3.36 ist in dieser Region nichts erkennbar. Auffällig ist dagegen ein Peak bei dem 5.8S rRNA Gen, obwohl laut Tabelle 3.12 die 5.8S rRNA an fünfter bzw. sechster Stelle steht. Dies liegt vermutlich an der vorherigen Filterung aller *Reads* < 15nt, die für Abbildung 3.36 verwendet wurden. Die 5.8S rRNA ist Teil eines kurzlebigen primären Transkriptes bestehend aus 17S, 5.8S und 26S rRNA, welches zu zwei *precursor* Molekülen prozessiert wird [Boesler et al., 2011]. Es können mehr als 43% der *Reads* aller Replikate annotierten Genen zugeordnet werden, von denen besonders viele für rRNA kodieren. Das heißt, dass kleine RNA für deren Regulation unabhängig von RrpC von besonderer Bedeutung sein könnten.

Wie bereits beschrieben ist die Menge der *Read*-Überlappungen mit dem Retrotransposon DIRS-1 im Gendeletionsstamm um 32% reduziert, was die experimentellen Daten von Stephan Wiegand bestätigen [Wiegand et al., Prep]. Die Reduzierung bzw. Erhöhung der Menge kleiner RNA lässt auf eine, durch das fehlende *rrpC* Gen verursachte, Abhängigkeit der RNA Produktion schließen. Die unregelmäßige Verteilung entlang des DIRS-1 Gens ist in Abbildung 3.34 zu sehen, neben den drei beobachteten Klassen kleiner RNA, die im *rrpC* Gendeletionsstamm konstant, halb oder mehr reduziert sind oder nur in diesem vorkommen. Die Verteilung auf dem *Sense* Strang und *Antisense* Strang ist dabei ebenfalls unregelmäßig. Das heißt, dass die mRNA vermutlich Primer-abhängig direkt als Vorlage für die RrpC dient. Die kleinen RNA, die auf beiden Strängen konstant sind, könnten primäre Produkte der Dicer ähnlichen Proteine sein. Das bedeutet, dass die kleinen RNA durch verschiedene RNAi-Mechanismen generiert werden [Hinas et al., 2007]. Da DIRS-1 Sequenzen in *Dictyostelium discoideum* wahrscheinlich einen Teil des Centromers ausmachen [Glöckner & Heidel, 2009], könnte die Reduktion kleiner RNA einen Effekt auf die chromosomale Stabilität besitzen, da die Mobilität des Retrotransposons weniger gehemmt wird [Slotkin & Martienssen, 2007, Wiegand et al., Prep]. Im Gegensatz dazu gibt es im Gendeletionsstamm eine 30-fach erhöhte *Read*-Überlappung mit Skipper, speziell in zwei bestimmten Regionen in denen ORF I und ORF II beginnen. Eine hochregulierte Anzahl kleiner RNA in Skipper wurde ebenfalls von Hinas *et al.* berichtet [Hinas et al., 2007]. Die Faltung der zwei auffälligen Regionen und die Faltung der Fusion beider Regionen (Abbildung 3.35) ergaben keine neuen oder bekannte miRNA. Es gab keine Ähnlichkeit zu den Einträgen der miRNA Datenbank und das miRNA Vorhersageprogramm miR-abela konnte keine mögliche *pre*-miRNA finden. Abbildung 3.35 zeigt, dass die Längenkriterien von 21-23 Basenpaaren nicht erfüllt werden. Die kleinen RNA, die im Gendeletionsstamm angereichert sind, könnten im Wildtyp einem Primer-abhängigen RNAi-Mechanismus dienen und deshalb reduziert sein [Wiegand et al., Prep]. Diese würden als Substrat durch RrpC zu doppelsträngiger RNA synthetisiert und anschließend durch RNAi degradiert werden [Wiegand et al., Prep]. Des Weiteren gibt es ebenfalls signifikante Hochregulierungen kleiner RNA in anderen getesteten *Repeats*.

Allgemein sind RdRP für die Kontrolle von Transposons nicht essentiell, wie mit Hilfe mathematischer Modelle gezeigt werden konnte, da alternative RNA basierte Strategien existieren [Crombach & Hogeweg, 2011]. Außerdem sind RdRP für das Überleben von *Dictyostelium discoideum* nicht essentiell, da Stephan Wiegand einen Gendeletionsstamm erstellen konnte, der lebensfähig war, was ebenfalls auf kompensatorische Effekte schließen lässt [Wiegand et al., Prep].

Insgesamt unterscheidet sich die Anzahl *Reads* in 2096 Regionen und 472 Genen zwischen Wildtyp und Gendeletionsstamm signifikant. Davon sind im Gendeletionsstamm die kleinen RNA in 186 Regionen reduziert und 1910 Regionen erhöht, was auch an den signifikanten Genunterschieden in Abbildung 3.38 zu sehen ist.

In einer anderen Mutationsstudie in *Arabidopsis thaliana* wurde nach dem Ausschalten der RdRP (Rdr2) eine Anreicherung kleiner RNA, besonders der miRNA und siRNA beobachtet [Lu et al., 2006].

Die Erzeugung kleiner RNA und die Regulierung von Transposons durch RrpA und RrpC geschieht auf unterschiedliche Weise [Wiegand et al., Prep]. Beim *rrpB* Gendeletionsstamm konnten keine signifikanten Änderungen der Anzahl kleiner RNA im Vergleich zum Wildtyp festgestellt werden [Wiegand et al., Prep].

5 Ausblick

In zukünftigen Arbeiten kann das Thema durch die stetig ansteigende Anzahl sequenzierter Genome immer weiter fortgeführt werden. Durch die Modifikation existierender und die Erstellung neuer Deskriptoren sind viele weitere Suchen möglich.

Zum Beispiel können die gefundenen Variationen des katalytischen Zentrums der *Hammerhead* Ribozym näher untersucht werden. Ein Ansatz ist das Erlauben von alternativen Nukleotidinsertionen vor und nach den Helices, wie z. B. beim *Chrysanthemum chlorotic mottle viroid* [de la Peña et al., 1999].

Da inzwischen bekannt ist, dass auch HHRz II existieren [Jimenez et al., 2011] könnte eine bisher nur auf wenige Organismen durchgeführte Suche auf allen Organismen ausgeführt werden. Das gleiche gilt für die Suche nach HDV Ribozymen.

Nach der experimentellen Auswertung und Analyse von Florentine Stix, Anne Kalweit und Susann Weißheit stellte sich heraus, dass der HHRz I Deskriptor (07.01.2011) eine zu kurze Helix I am 5' 3' Ende besitzt und somit notwendige Tertiärinteraktionen zum Teil nicht gebildet werden können. Diese Sequenzen sind FP und sollten in einer erneuten Suche mit verlängertem *Stem* Ia ausgeschlossen werden. Außerdem schlug Florentine Stix in ihrer Bachelorarbeit vor, einen *Bulge* auszuschließen und einen *internal Loop* I zu definieren.

Die bisherige Suche nach *Hairpin* Ribozymen umfasst nur einen Teil des Motivs, so dass aufwendig manuell untersucht werden muss, ob weitere Teile des Motivs vorhanden sind. Dazu müssten im nächsten Schritt die Sequenzen (17.03.2010) in 5' und 3' Richtung erweitert werden. Dies könnte jedoch durch eine Beschreibung des minimalen Motivs und eine erneute Suche vereinfacht werden. Des Weiteren sollten die $\Delta\Delta G$ Filtereinstellungen angepasst werden, so dass die TP aus subviral DB nicht entfernt werden.

Bezüglich RNAhit könnten optionale Änderungen der Faltungsprogramme eingebaut werden, um die Struktur z. B. bei einer Organismus typischen Temperatur falten zu lassen und somit eine bessere Abschätzung der freien Energie zu erlangen. Die Motivtreffer werden gemäß einer Motivbeschreibung als Punkt-Klammer-Struktur ausgegeben. Der anschließende Vergleich der Faltungen basiert zur Zeit ausschließlich auf dem Vergleich der minimalen freien Energie und könnte durch den zusätzlichen Vergleich der Punkt-Klammer-Strukturen verbessert werden, um alternative FP Strukturen mit gleichem ΔG Wert ausschließen zu können. Des Weiteren könnten dem Benutzer alternative Such- und Faltungsprogramme angeboten werden. So wäre als nächstes Faltungsprogramm RNAfold hinzugefügt worden oder ein Faltungsprogramm, das Pseudoknoten berücksichtigen kann. Ein neuer objekt-orientierter Ansatz des Programmes RNAhit 2.0.0, der diese Erweiterungen durch einen besseren modularen Aufbau vereinfachen könnte, wurde begonnen, aber nicht fertig gestellt und entspricht nicht den Funktionen der aktuellen Version.

Zukünftig könnten weitere Filteroptionen hinzugefügt werden, um FP Treffer zu reduzieren, wie z. B. ein Filter für Tertiärinteraktionen. Dazu müssen jedoch weitere katalytisch aktive Motive untersucht und besser verstanden werden, da diese hoch variabel sind und die Gefahr besteht, bei zu strikter Motivdefinition aktive Motive zu entfernen.

Die bereits gefundenen Treffer können durch hinzukommende Annotationen besser ausgewertet und durch weitere Experimente validiert werden. Von besonderem Interesse ist z. B. die Überprüfung der katalytischen Aktivität der HHRz III Variation aus *Arabidopsis lyrata*.

Im zweiten Projekt sollten die *Read*-Lokalisierungen und die signifikanten Unterschiede, besonders bei den ermittelten Genen, näher untersucht werden. Reads mit einer Länge von 21 nt könnten mit einem miRNA Vorhersageprogramm durchsucht werden und somit bekannte und neue miRNA identifizieren.

6 Zusammenfassung

Es gibt mehrere Möglichkeiten mit unterschiedlicher Komplexität, dass eine RNA-Sequenz eine endonukleolytische Spaltung bzw. Ligation des Phosphodiesterrückgrates der eigenen Sequenz durchführt [Cochrane & Strobel, 2008, Fedor, 2009]. Diese RNA-Motive mit einer katalytischen Aktivität werden als Ribozyme bezeichnet und sind mögliche Reste einer RNA-Welt ohne Proteine. Es existieren große (> 200 nt) und kleine (50-150 nt) Strukturen, wie die *Hammerhead*- und *Hairpin* Ribozyme, HDV ähnliche Ribozyme und das VS Ribozym. In den letzten Jahren haben verschiedene Studien gezeigt, dass katalytische RNA-Motive trotz geringer Wahrscheinlichkeit weiter verbreitet sind als bisher angenommen. So wurden HDV ähnliche Ribozyme im Menschen [Salehi-Ashtiani et al., 2006] und zahlreichen anderen Organismen gefunden [Webb et al., 2009], wo sie z. B. der Spaltung von Retrotransposons dienen [Ruminski et al., 2011]. *Hammerhead* Ribozyme (Typ I, II, III) kommen in allen Domänen des Lebens vor [de la Peña & Garcia-Robles, 2010b, Seehafer et al., 2011, Perreault et al., 2011]. Bekannte natürliche *Hairpin* Ribozyme wurden z. B. in viraler Satelliten RNA gefunden [Rocheleau & Pelchat, 2006] und nach wie vor ist das VS Ribozym das Einzige bekannte dieser Struktur [Saville & Collins, 1990]. Von besonderem Interesse in dieser Arbeit sind *Hammerhead* Ribozyme. Sie besitzen ein konserviertes katalytisches Zentrum, das von drei variablen Helices umgeben ist. HHRz benötigen für eine katalytische Aktivität unter physiologischen Bedingungen Magnesium Ionen und die Interaktion peripherer Elemente in der Tertiärstruktur. Sie wurden erstmals in Viroiden [Hutchins et al., 1986] und in Satelliten RNA des *Tobacco ringspot Virus* [Prody et al., 1986] entdeckt und dienen dort der Spaltung von multimeren Fragmenten zu Monomeren während der *Rolling circle* Replikation. Das Problem der HHRz-Suche bzw. allgemein der RNA-Motivsuche besteht darin, dass die Konservierung der Struktur größer ist als die der Sequenz und dass eine korrekte Sequenz nicht zwingend die gesuchte Struktur einnimmt, die für eine katalytische Aktivität benötigt wird.

Dieses Problem sollte durch einen geeigneten Filter gelöst werden. Angesichts der signifikant gestiegenen Sequenzverfügbarkeit galt es zunächst, in öffentlichen Datenbanken nach neuen Beispielen bekannter RNA-Motive zu suchen, um die Verbreitung, Variation und Evolution dieser Motive zu untersuchen. Ein strukturbasierter Ansatz zeigte sich bereits in früheren Studien als nützlich [Przybilski et al., 2005, Webb et al., 2009]. Aus den primären Suchergebnissen sollten anschließend durch eine *Pipeline* die Sequenzen ermittelt werden, die eine mögliche katalytische Aktivität besitzen. Die resultierenden Daten sollten daraufhin analysiert, annotiert und visualisiert werden, um mögliche Rückschlüsse über die Funktion der Motive zu erhalten.

Es wurden zahlreiche RNA-Motive mit Hilfe strukturbasierter Deskriptoren erstellt, mit verschiedenen Suchprogrammen gesucht und unterschiedliche Faltungs- und Filterparameter getestet. Der Hauptfilterschritt bestand aus dem Vergleich freier Energien. Dazu wurde die minimale freie Energie der gefundenen Sequenz berechnet. Anschließend wurde dieselbe Sequenz in das gesuchte Motiv gezwungen und erneut die freie Energie bestimmt. Der Vergleich und die darauffolgenden Filterschritte sind im Programm RNAhit implementiert. Es ruft unter den gesetzten Einstellungen die einzelnen Unterprogramme auf, leitet deren Ergebnisse weiter und filtert sie. Die Einstellungen wurden durch Optimierungsverfahren ermittelt bzw. mit Hilfe experimenteller Daten angepasst. Die Sequenzen, bei denen das Motiv zutrifft, wurden korrekt ausgegeben, was bedeutet, dass die angewandten Filterschritte für die Trennung zwischen TP und FP geeignet sind. Die Ergebnisse wurden im Dateisystem in verschiedenen Formaten und in einer MySQL-Datenbank abgelegt und über einen modifizierten ProServer sowie mit Hilfe der Ensembl Webseite visualisiert. Es wurden 76 mögliche *Hairpin* Ribozyme in 42 Eukaryonten und 2 Bakterien identifiziert, die einen Teil des Motivs erfüllen, bei denen jedoch die fehlenden Segmente überprüft werden müssen. Die sonst geringe Menge im Vergleich zu den HHRz deutet auf eine unabhängige Verbreitung. Die Kombination von *Hairpin* Ribozymen und HHRz könnte spezifisch für sub-virale Pathogene sein und das Vorkommen in diesen weist auf einen frühen Ursprung. Des Weiteren zeigte die verallgemeinerte Suche eines HDV Ribozym ähnlichen Deskriptors in *Dictyostelium discoideum* mehr als 1750 Treffer, von denen 413 mit Genen assoziiert werden. Neue Beispiele VS Ribozym ähnlicher Strukturen konnten nicht gefunden werden. Von allen gesuchten RNA-Motiven kommt das *Hammerhead* Ribozym, besonders Typ I, am häufigsten in den Organismen vor. Die Suche nach Typ I, II und III zeigte in repetitiven Regionen des Krallenfrosches *Xenopus tropicalis* einen modularen Aufbau, in dem alle drei Typen in einer Region möglich sind. In einer genomweiten Suche gelang es mit dem Programm RNAhit aus der enormen Datenmenge von mehr als 144 Giga Byte (GB) neue *Hammerhead* Ribozyme zu identifizieren, welche in der Publikation [Seehafer et al., 2011] veröffentlicht wurden. Die Abschätzung der HHRz III Motivwahrscheinlichkeit zeigte eine 10-fach größere Trefferanzahl, als durch Zufall erwartet. Aus mehr als 60000 primären Treffern aus Eukaryonten (95%), Bakterien (2%) und sub-viralen Pathogenen (3%) konnten 284 einzigartige HHRz bestimmt werden, von denen 122 bereits bekannt waren. Die anderen 160 neuen HHRz aus 50 Eukaryonten und 3 Bakterien zeigen eine große Heterogenität in ihrer Sequenz und Sequenzlänge, welche das Erstellen eines *Multiple Sequence Alignments* erschwerte.

Die meisten Treffer stammen aus *Hydra magnipapillata* (35) und *Schistosoma mansoni* (24). Allgemein befinden sich die HHRz I und III Treffer in repetitiven Sequenzbereichen, in Introns proteinkodierender Gene oder in intergenischen Regionen. Einige dieser intergenischen Regionen stehen im Zusammenhang mit transposablen Elementen. Die Sequenzähnlichkeiten der HHRz zwischen den Organismen könnten sich durch mobile Elemente, wie Retrotransposons oder sub-virale Pathogene verbreitet haben. Innerhalb der Organismen wurde zum Teil nur ein HHRz III Motiv gefunden, welches jedoch mehrmals vorkommen kann. Dies könnte ein Anzeichen für alternative Variationen oder Ribozyme sein. Die Homologiesuche eines neu gefundenen HHRz III in *Arabidopsis thaliana* zeigte eine weitere mögliche Variation des HHRz in *Arabidopsis lyrata*. Des Weiteren wurden Erweiterungen des *Loop* I in HHRz III häufiger gefunden als Strukturen mit erweitertem *Loop* II, was aufgrund der Substratlänge für eine Spaltung in *trans* zu erwarten war. In Zusammenarbeit mit Franziska Hoffgaard konnten zwischen und innerhalb der Helices, je nach HHRz-Typ, bisher unbekannte signifikante co-evolutionäre Zusammenhänge gezeigt werden. Die größere Übereinstimmung co-evolvierender Positionen der HHRz I mit sub-viralen Sequenzen im Vergleich zu HHRz III könnte auf ein höheres Alter deuten, vorausgesetzt sie besitzen einen gemeinsamen Ursprung. Die häufigsten beobachteten HHRz I Längen liegen zwischen 50-60 nt mit variierenden *Loop*-Größen zwischen 5, 7 oder 9 nt, wobei die Motivlängen unabhängig von den Genomgrößen sind. Durch zum Teil manuelle Korrektur des MSA der 160 neu identifizierten HHRz III konnte ein phylogenetischer Baum erzeugt werden. Dieser Baum zeigt verschiedene Cluster von HHRz, die den einzelnen Organismen zugeordnet werden können, was bedeutet, dass innerhalb der Organismen die Motive konserviert sind. Aber auch zwischen den Organismen sind zum Teil Gattungen, wie innerhalb der *Drosophila* und Klassen, wie innerhalb der Vögel, erkennbar. Die Verallgemeinerung des katalytischen Zentrums an den Positionen 3 und 8 im HHRz III zeigte als häufigste beobachtete Kombination U3A8. Obwohl U3A8 im Vergleich zu C3G8 eine verminderte selbst-spaltende Aktivität besitzt. Dies könnte auf ein höheres Alter innerhalb der HHRz III deuten, falls sie einen gemeinsamen Ursprung besitzen. Diese und andere Variationen sprechen für eine Entwicklung des Motivs und Optimierung der Spaltungsrate in der Evolution. Allgemein könnten beobachtete Variationen eine Anpassung an Umweltbedingungen sein. Das Vorkommen in sub-viralen Pathogenen und den unterschiedlichsten Organismen des Lebens ist ein Hinweis für eine sehr frühe Evolution. HHRz könnten mehrere Ursprünge haben, da die Sequenzen zwischen den Organismen sehr heterogen und ihre Fundorte sehr variabel sind. Außerdem konnte in SELEX Experimenten unter annähernd physiologischen Bedingungen gezeigt werden, dass aus einer Zufallssequenz heraus ein HHRz entstehen kann [Salehi-Ashtiani & Szostak, 2001]. Die Entwicklung könnte unabhängig voneinander erfolgt sein. Das HHRz-Motiv ist im Vergleich zu den anderen Ribozymen die einfachste Lösung für das biologische Problem der Spaltung einer Nukleinsäurekette ohne Proteine. Über die mögliche Funktion der Ribozyme lässt sich zusammenfassen, dass alle die Teilung einer transkribierten RNA durch das Motiv gemeinsam haben und dadurch eine Sequenz getrennt bzw. verbunden werden kann. Somit kann dies für die Hemmung, Prozessierung oder die Integration der RNA von Bedeutung und für weitere Interaktionen eine Voraussetzung sein.

In einem zweiten Projekt dieser Arbeit galt es RNA *Deep Sequencing* Daten eines *Dictyostelium discoideum* RdRP Wildtyp und Gendeletionsstamms, jeweils in zwei Replikaten, auszuwerten.

Viele Eukaryonten verwenden im RNAi Mechanismus eine RdRP, die eine dsRNA synthetisiert, aus der kleine RNA produziert werden, die zum Abbau einer Ziel-RNA führen [Maida & Masutomi, 2011]. Das in *Dictyostelium discoideum* für die RdRP kodierende Gen *rrpC* wurde durch homologe Rekombination entfernt [Wiegand et al., 2011] und dessen Folgen auf die Produktion kleiner RNA näher untersucht. Aus vorherigen Experimenten war bekannt, dass die mRNA Menge des Retrotransposons DIRS-1 im *rrpC* Gendeletionsstamm erhöht ist [Kuhlmann et al., 2005]. Außerdem konnte in Northern Blot Experimenten im *rrpC* Gendeletionsstamm eine große Reduzierung kleiner RNA für DIRS-1 beobachtet werden [Wiegand et al., Prep]. Dies konnte mit den Sequenzierungsdaten bestätigt werden, die eine Reduktion kleiner RNA um 42% zeigten. Daraus folgt, dass RrpC für das transkriptionelle *Silencing* von DIRS-1 mRNA verantwortlich ist.

Ein erster Schritt in der Auswertung der Daten war die Erstellung eines Qualitätsberichtes, der zeigte, dass sich in einem Großteil der *Reads* Reste einer für die Sequenzierung benötigten Adaptersequenz befinden. Diese wurden entfernt und führten zu einer Längenverteilung der *Reads*. Die größte Häufigkeit besaßen *Reads* mit einer Länge von 21 nt, von denen ein Großteil DIRS-1 zugeordnet werden konnte und die siRNA oder miRNA sein könnten. Die *Reads* enthielten eine hohe Anzahl an Duplikationen und stammen zum Teil aus repetitiven Regionen mit wenig Variabilität. Die bearbeiteten *Reads* wurden gegen das *Dictyostelium discoideum* Genom und gegen eine *Repeat*-Bibliothek abgebildet. Die Auswertung zeigte, wie bereits erwähnt, eine Reduktion aller kleinen RNA im *rrpC* Gendeletionsstamm, besonders für das LTR-Retrotransposon DIRS-1. In *Dictyostelium discoideum* ist DIRS-1 das am häufigsten vorkommenden, mobile Element [Eichinger et al., 2005]. Da DIRS-1 Sequenzen ein Teil des Centromers ausmachen [Glöckner & Heidel, 2009], könnte die Reduktion kleiner RNA einen Effekt auf die chromosomale Stabilität besitzen, da die Mobilität des Transposons weniger gehemmt wird. Die *Read*-Verteilung, speziell bei DIRS-1, zeigte drei Gruppen kleiner RNA, die konstant, abhängig und unabhängig von RrpC sind. Das zweit häufigste mobile Element in *Dictyostelium discoideum* ist Skipper [Eichinger et al., 2005], welches im Gendeletionsstamm eine signifikant größere Anzahl kleiner RNA besaß als im Wildtyp. Diese konnten speziell zwei auffälligen Regionen zugeordnet werden.

Die Hochregulierung kleiner RNA im Gendeletionsstamm deutet im Wildtyp auf einen Primer-abhängigen RNAi-Mechanismus. Diese würden als Substrat durch RrpC zu doppelsträngiger RNA synthetisiert und anschließend durch RNAi degradiert werden [Wiegand et al., Prep].

Im Wildtyp wurden die meisten *Reads* Chromosom 1 zugeordnet, gefolgt von dem extra-chromosomalen rDNA-Palindrom. Im *rrpC* Gendeletionsstamm ist es umgekehrt. Das *Mapping* zeigte zwei mögliche Regionen auf dem extra-chromosomalen rDNA Palindrom in denen die *Reads* aligniert wurden. Diese gleichen sich zu 99%. Mehr als 43% aller *Reads* alignieren zu annotierten Genen, von denen ein Großteil für ribosomale Komponente kodiert. Das heißt, dass kleine RNA für deren Regulation unabhängig von RrpC von besonderer Bedeutung sind.

Insgesamt unterscheiden sich die *Reads* von 2096 Regionen und 472 Genen zwischen Wildtyp und Gendeletionsstamm signifikant. Davon sind im Gendeletionsstamm die kleinen RNA in 186 Regionen reduziert und in 1910 Regionen erhöht. Es konnte in *Dictyostelium discoideum* schließlich ein Einfluss durch RrpC auf Retrotransposons bestätigt und für DIRS-1 und Skipper verschiedene Mechanismen vorgeschlagen werden.

Summary

There are various options with varying complexity for an RNA sequence to self-cleave or self-ligate the phosphodiester bond of its own sequence [Cochrane & Strobel, 2008, Fedor, 2009]. These RNA motifs with catalytic activity are called ribozymes and are potential relics of an ancestral RNA world in absence of proteins. Their length distribution differs from 50 nt to more than 200 nt. Ribozymes are grouped into Hammerhead-, Hairpin-, HDV and VS ribozymes. Studies of the past years showed, that despite of the low probability, catalytic RNA motifs are widespread in nature. For example HDV like ribozymes were found in human [Salehi-Ashtiani et al., 2006] and numerous other organisms [Webb et al., 2009], where they are used to cleave retrotransposable elements [Ruminski et al., 2011]. Hammerhead ribozymes of type I, II and III could be identified in all domains of life [de la Peña & Garcia-Robles, 2010b, Seehafer et al., 2011, Perreault et al., 2011]. Known natural hairpin ribozymes were found in viral satellite RNA [Rocheleau & Pelchat, 2006] and the *Neurospora* varkud satellite ribozyme is the only known of this structure [Saville & Collins, 1990]. This thesis investigates especially hammerhead ribozymes, that consist of a conserved core forming a three way junction, which is surrounded by three variable helices. HHRz require a physiological Mg^{2+} concentration and a tertiary interaction of peripheral elements to self cleave under physiological conditions. The HHRz motif was discovered in viroids [Hutchins et al., 1986] and in satellite RNA of the *tobacco ringspot virus* [Prody et al., 1986], where it is used during the rolling circle replication, to cleave multimeric fragments into monomers. The problem of searching HHRz or in general of searching RNA motifs is the level of conservation. Structures are more conserved than sequences and a correct sequence match does not mean that the sequence folds into the searched structure, which is required for the catalytic activity.

The purpose of this thesis was to solve this problem by using appropriate filters. In the light of the significant increase of available sequences in public databases, new examples of known RNA motifs were searched to explore the distribution, variation and evolution of those motifs. Therefore a structure based approach was chosen, which was shown to be useful in earlier studies [Przybilski et al., 2005, Webb et al., 2009]. A pipeline should then detect in the large amount of primary hits those sequences with a potential catalytic activity. The resulting data should then be analyzed, annotated and visualized to draw conclusions about the function of the motifs.

Numerous structure based descriptors were designed and searched using diverse search programs. The result was folded using different folding programs and afterwards filtered with variable filter parameters. The main filter step was the comparison between the minimum free energy of a sequence with the free energy of that sequence forced into a given structure. The main program of this thesis, next to many other, is called RNAhit. It uses the given settings to call the different sub-routines and forward their results, which thereafter are filtered. The settings were determined using optimization techniques and experimental data. The correctness of the filter steps could be shown by the accurate separation of catalytically active (true positives) and inactive (false positives) known motifs. The results were stored in the file system in various formats and in a MySQL database. The data were visualized by a modified ProServer and by using the Ensembl website. The search for hairpin ribozymes showed 76 hits in 42 eucaryotes and 2 bacteria, but only for a part of the motif. The missing parts should be checked in a second review. The small amount of hits in comparison to the number of HHRz indicates an independent distribution of the hairpin ribozyme. The combination of both ribozymes could be specific for sub-viral pathogens. The occurrence in sub-viral pathogens is an evidence for an early origin. Furthermore the generalized search in *Dictyostelium discoideum* for HDV like ribozymes revealed more than 1750 hits including 413 hits, that can be associated with genes. New examples of varkud satellite like ribozymes could not be found. Finally all searches in different organisms resulted in the observation, that hammerhead ribozymes are the most frequent known ribozymes, especially of type I. The search for type I, II and III showed a modular organization in repetitive regions of *Xenopus tropicalis*, where all three types could be found in one region. In the background of the enormous amount of data (144 GB) it was possible to filter, identify and publish new putative hammerhead ribozymes using the program RNAhit [Seehafer et al., 2011]. The number of primary hits was 10 fold higher than expected by chance. More than 60000 primary hits from eucaryotes (95%), bacteria (2%) and sub-viral pathogens (3%) matched to the motif description and were filtered using thermodynamic parameters. This resulted in 284 unique HHRz including 122 known sequences. The other 160 new HHRz from 50 eucaryotes and 3 bacteria were further analyzed and displayed a high heterogeneity in their sequence and sequence length, which complicated the creation of a multiple sequence alignment. The highest number of unique hits could be determined in *Hydra magnipapillata* (35) and *Schistosoma mansoni* (24). The hits were located in repetitive sequence regions, in introns of protein coding genes and in intergenic regions. Some of these intergenic regions are linked to transposable elements. These and other mobile elements could be used to spread the motif. In some cases the whole genome contained only one HHRz III motif, but in multiple copies. This could be an evidence for alternative variations or ribozymes. Homology searches of the new found HHRz sequences in *Arabidopsis thaliana* led to another possible variation in *Arabidopsis lyrata*. In addition both loops were extended to find HHRz III motifs that are capable to cleave in *trans*. Due to the length of the substrate the higher amount of hits was expected for motifs with a loop I extension. This was the case, despite of the same motif probability in random sequences.

Furthermore there is a so far unknown co-evolution of sequence positions between and within helices of the HHRz, depending on the type. This observation is based on the mutual information calculated by Franziska Hoffgaard. The high similarity of co-evolutionary positions of HHRz I to sub-viral sequences in comparison to HHRz III points to a higher age of type I, assuming a common origin. The most frequent observed length of the HHRz I is between 50 to 60 nt with varying loop sizes between 5, 7 or 9 nt. There is no correlation between the motif length and the genome size. A phylogenetic tree could be created with help of the partially manual corrected multiple sequence alignment of the new 160 HHRz III. This tree clusters different HHRz into single organisms, which could mean, that the motifs are conserved within the organisms. On the other hand there are also clusters of species like the genus *Drosophila* and even clusters, that can be grouped into the class of birds. The generalization of the HHRz III in the catalytic center showed that in position 3 and 8 the most frequent observed combination is U3A8, although U3A8 has a reduced self-cleaving activity in comparison to C3G8. This could be an evidence for a higher age within the HHRz type III, if they have a common origin. This and other variations are an indication for the development of the motifs and an optimization of the cleavage rate during evolution. In general observed variations could be an adaption to environmental conditions. The occurrence in sub-viral pathogens and different domains of life is an evidence for an early evolution. HHRz could have more than one origin, due to the heterogeneity of the sequences between the organisms and the variability of the identified locations. Another evidence for that is an observation made by Salehi *et al*. They could show in a SELEX experiment under physiological conditions, that catalytically active HHRz can occur in random sequences [Salehi-Ashtiani & Szostak, 2001]. Thus the motifs probably developed independent of each other. The HHRz is the simplest solution to cleave a nucleic acid without proteins in comparison to the other ribozymes . The function of all ribozymes can be summarized in the point that all ribozymes are responsible for the cleavage or ligation of a transcribed RNA. In this way it could be important for the prohibition, processing or integration of RNA and it could be a requirement for further interactions.

The second project of the thesis was the analysis of deep sequencing data of the model organism *Dictyostelium discoideum*. Therefore two replicates of an RdRP wild type and an *rrpC* gene deletion strain were analyzed.

Many eucaryotes utilize RdRP in the RNAi mechanism to synthesize double-stranded RNA, which is processed into small RNA, that lead to degradation of a target RNA [Maida & Masutomi, 2011]. There are three genes (*rrpA*, *rrpB*, *rrpC*), that encode RdRP in *Dictyostelium discoideum*. The gene *rrpC* was removed by homologous recombination [Wiegand et al., 2011] and the effects on the production of small RNA were analyzed. It is known from previous experiments, that in *rrpC* gene deletion strains the mRNA expression of the retrotransposable element DIRS-1 is increased [Kuhlmann et al., 2005]. In addition Northern Blots showed for DIRS-1 a significant reduction of small RNA in the *rrpC* knockout [Wiegand et al., Prep]. The sequencing data confirmed these results and showed a reduction of 42% in comparison to the wild type. This means, that RrpC is responsible for the silencing of DIRS-1 mRNA.

The first step in the analysis of the sequenced small RNA was the generation of a quality report. The report referred to many reads that contain parts of an adapter sequence, which is used during sequencing. This part was removed and led to a length distribution of the reads. The most frequent length was 21 nt and the reads could be related to DIRS-1. These sequences could be siRNA or miRNA. The reads included a high number of duplications, which could be mapped to repetitive regions with a low variability. The processed reads were mapped against the genome of *Dictyostelium discoideum* and against a repeat library. As already mentioned the results showed a reduction of small RNA in the *rrpC* gene deletion strain, especially for the retrotransposon DIRS-1. DIRS-1 is in *Dictyostelium discoideum* the most frequent mobile element [Eichinger et al., 2005]. The sequences are part of the centromers [Glöckner & Heidel, 2009] and the reduced amount of small RNA could have effects on the chromosomal stability, because of the reduced inhibition of the mobility. The asymmetrical distribution of the reads, that map to DIRS-1 showed clusters of three groups. Those that have a constant distribution in all replicates, those that depend on the RdRP and those that are independent. The second most frequent mobile element in *Dictyostelium discoideum* is Skipper [Eichinger et al., 2005], which had a significant higher amount of small RNA in the knockout than in the wild type. The significant increase of small RNA in Skipper showed two conspicuous regions. The up regulation of small RNA in the gene deletion strain is an evidence for a primer-dependent RNAi mechanism in the wild type. Small RNA would serve as substrate, which is elongated and synthesized to a double-stranded RNA by RrpC, that afterwards is degraded by RNAi [Wiegand et al., Prep].

The most reads of the wild type were mapped to chromosome 1 followed by the extra-chromosomal rDNA palindrome. In the *rrpC* knockout it was the other way around. The extra-chromosomal rDNA palindrome encodes for ribosomal components. These are annotated at the 5' end and have a 99% identity with the complement strand at the 3' end. The reads were mapped to these two regions. In general 43% of the reads in all replicates could be aligned to annotated genes. The most of them were aligned to ribosomal components, which suggests a particular meaning for the regulation of these components by small RNA.

The reads differ between wild type and gene deletion strain in 2096 regions and 472 genes significantly. The amount of produced small RNA by the RdRP Knockout is down regulated for 186 and up regulated for 1910 regions.

In summary an influence of RrpC on retrotransposable elements could be confirmed and two distinct mechanisms for DIRS-1 and Skipper were proposed.

Glossar

CONSENSUS

Ist ein Motivsuchprogramm, dass mehrere paarweise Ähnlichkeiten bildet und so Signale identifiziert. [Hertz & Stormo, 1999]

Dicer

Dicer sind Enzyme, die mögliche kleine *guide* RNA produzieren durch die spezifische Spaltung doppelsträngiger RNA [Bernstein et al., 2001].

Distributed Annotation System

Ist ein Protokoll unter Verwendung von XML und HTTP, um zahlreiche biologische Daten aus verschiedenen Quellen dynamisch zu integrieren [Finn et al., 2007].

E-Value

Der *E-Value* einer Datenbanksuche gibt an, wie hoch die Anzahl der *Alignments* mit gleichem oder besserem Score ist, die durch Zufall zu erwarten wären. Je kleiner der *E-Value* desto signifikanter ist der Score. Ein *E-Value* der Größe 1 entspricht in einer Datenbank der gleichen Größe einem Treffer mit gleichem Score allein durch Zufall. Weitere Informationen sind unter http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html zu finden.

Ensembl

Ist ein gemeinsames Projekt des EBI und Wellcome Trust Sanger Institute (WTSI) zur Annotation, Analyse und Darstellung von Genomdaten.

EST

Sind kurze Sequenzen, die der Katalogisierung von mRNA dienen [Hüttenhofer et al., 2001].

FPKM

Der FPKM Wert entspricht der relativen Fragmentmenge pro Kilobasen eines Exon Models pro 1 Million alignierter Fragmente.

Hash

Ein *Hash* ist ein Schlüssel-Wert Paar, bei dem ein einzigartiger Schlüssel einem Wert zugeordnet wird. Durch den Aufruf des Schlüssels wird der Wert abgefragt. Bei zwei gleichen Schlüsseln wird der alte Wert durch den neuen überschrieben.

Host

Ein Host ist in der Informatik ein Server.

Kalign

Kalign ist ein schnelles und genaues MSA Programm, geeignet für große Datenmengen.
[Lassmann & Sonnhammer, 2005]

Levenshtein-Distanz

Die Levenshtein-Distanz, als Maß der Ähnlichkeit zweier Sequenzen, ist die minimale Anzahl von Insertionen, Deletionen und Substitutionen, um eine Sequenz in eine andere umzuwandeln. [Levenshtein, 1966]

MI

Die *Mutual Information* $MI_{ij} := \sum_{\sigma_i \in S} \sum_{\sigma_j \in S} P(\sigma_i, \sigma_j) * log_2 \frac{P(\sigma_i, \sigma_j)}{P(\sigma_i)P(\sigma_j)}$ ist ein Maß aus der Informationstheorie [Shannon, 1948]. Sie bestimmt die Informationsmenge einer Sequenzposition gegenüber einer anderen Position. Ein co-evolutionärer Zusammenhang durch einen unbekannten selektiven Druck besteht, wenn zwischen diesen beiden Positionen bzw. Spalten eines MSA Korrelationen existieren, die durch Basenaustausch der Positionen erkennbar werden [Hoffgaard et al., Prep].

MUSCLE

MUSCLE steht für '***MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation*' und ist, wie ClustalW, ein heuristisches, progressives MSA Programm. MUSCLE erreicht jedoch eine bessere durchschnittliche Genauigkeit und ist schneller als ClustalW, je nach gewählten Optionen. [Edgar, 2004]

NCBI

NCBI ist Teil des US National Institues of Health und wurde 1988 gegründet. Sie bieten Genomsequenzierungsdaten, biomedizinische Artikel und zahlreiche weitere biotechnologisch relevante Informationen an. Weitere Informationen unter http://www.ncbi.nlm.nih.gov/

P-Value

Der *P-Value* gibt an, wie wahrscheinlich es ist den Datenpunkt zu erhalten, wenn kein Unterschied existiert. Ein *P-Value* der Größe 0.01 entspricht also einer Wahrscheinlichkeit von 1%, dass der Datenpunkt falsch positiv ist.

Patches

Patches sind Aktualisierungen, um Verbindungen zwischen alten Versionen der genomischen DNA des ersten *Assembly* zur aktuellen Version herzustellen. Dabei wird zwischen neuen *Patches*, die neuen allelischen Loci entsprechen und fix *Patches* unterschieden, die falsche *Assembly* korrigieren. Weitere Informationen befinden sich unter http://www.ensembl.org/.

Phred Quality Score

Der Phred Quality Score ist ein Basen spezifischer Score des Programmes PHRED, der logarithmisch im Zusammenhang mit dem wahrscheinlichen Fehler steht [Richterich, 1998].

Protein Datenbank

PDB ist eine Datenbank aus einer internationalen Zusammenarbeit (USA, Europa, Japan) zur Archivierung biologischer, makromolekularer Strukturen, Koordinaten und Informationen, die kristallografisch über Elektronenmikroskopie oder Kernspinresonanzspektroskopie gewonnen wurden [Berman et al., 2007].

Pseudoknoten

Pseudoknoten sind RNA Struktur, die aus zwei Helices bestehen und durch Einzelstränge miteinander verbunden sind [Staple & Butcher, 2005] (siehe am Beispiel des HDV Ribozymes).

Ribose Zipper

In der RNA-Tertiärstruktur wird die Ausbildung von Wasserstoffbrückenbindungen zwischen den Ribosen verschiedener Regionen einer oder zweier RNA-Ketten als Ribose Zipper bezeichnet [Tamura & Holbrook, 2002].

satelliten DNA

Satelliten DNA (Tandem *Repeats*) sind hintereinander gereihte, sich wiederholende Abschnitte auf der DNA. [Cegan et al., 2012]

Spaltungsrate

Um die Spaltungsrate berechnen zu können muss zunächst der Anteil gespaltener RNA zu einem Zeitpunkt t anhand der Pixel des PhosphorImager Bildes bestimmt werden $F_t = \frac{Produkt_1 + Produkt_2}{Substrat + Produkt_1 + Produkt_2}$ [Kalweit et al., 2011]. Dieser Anteil wird anschließend für jeden Zeitpunkt berechnet. Die Spaltungsrate k_{obs} ergibt sich durch das Fitten der Daten an die Gleichung $F_t = F_0 + F_\infty(1 - e^{-k*t})$ [Stage-Zimmermann & Uhlenbeck, 1998], wobei t der Zeit-, F_0 der Start- und F_∞ dem Endpunkt der Spaltungsreaktion entsprechen.

UPGMA

UPGMA ist die Abkürzung für '*Unweighted Pair Group Method with Arithmetic mean*' und ist eine hierarchische Clustering Methode [Sokal & Michener, 1958], die unter der Annahme einer molekularen Uhr (alle Taxa evolvieren gleich schnell) einen phylogenetischen Baum generieren kann.

Version

Bei der Versionsnummerierung werden drei Zahlen unterschieden. Die erste Zahl wird bei fundamentalen Änderungen erhöht, wie z. B. bei einem Neuanfang oder einer grundlegenden Umstrukturierung des Programmes. Die zweite Zahl steht für die funktionale Erweiterung und neue Eigenschaften des Programmes. Die dritte Zahl zeigt Verbesserungen von Programmfehlern an.

WINNOWER

Ist ein Graph-basierter, iterativer Algorithmus, der Motive durch das Entfernen von Hintergrundrauschen findet. Dabei werden Kanten in Gruppen zusammengefasst und nicht zuordbare Kanten aus dem Graph gelöscht, die nicht zum Signal gehören. [Pevzner & Sze, 2000]

Literaturverzeichnis

[Altschul et al., 1990] Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3), 403–410. PMID:2231712. 4.1.1

[Altschul et al., 1997] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25, 3389–3402. PMID:9254694. 2.4.12, 2.4.17

[Anders & Huber, 2010] Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11, R106. 2.4.18

[Aravin et al., 2008] Aravin, A., Sachidanandam, R., Bourchis, D., Schaefer, C., Pezic, D., Toth, K., Bestor, T., & Hannon, G. (2008). A pirna pathway primed by individual transposons is linked to de novo dna methylation in mice. *Mol Cell*, 31(6), 785–799. PMID:18922463. 1.3.2

[Athavale et al., 2012] Athavale, S., Petrov, A., Hsiao, C., Watkins, D., Prickett, C., Gossett, J., Lie, L., Bowman, J., O'Neill, E., Bernier, C., Hud, N., Wartell, R., Harvey, S., & Williams, L. (2012). Rna folding and catalysis mediated by iron (ii). *PLoS One*, 7(5), e38024. PMID:22701543. 1.3.4

[Babak et al., 2005] Babak, T., Blencowe, B., & Hughes, T. (2005). A systematic search for new mammalian noncoding rnas indicates little conserved intergenic transcription. *BMC Genomics*, 6, 104. PMID:16083503. 1.4

[Bachellerie et al., 2002] Bachellerie, J., Cavaille, J., & Hüttenhofer, A. (2002). The expanding snorna world. *Biochemie*, 84, 775–790. PMID:12457565. 1.3.2

[Bailey & Elkan, 1995] Bailey, T. & Elkan, C. (1995). The value of prior knowledge in discovering motifs with meme. *Syst Molecular Biol*, 3, 21–29. PMID:7584439. 1.5

[Bajaj et al., 2011] Bajaj, P., Steger, G., & Hammann, C. (2011). Sequence elements outside the catalytic core of natural hairpin ribozymes modulate the reactions differentially. *Biol Chem*, 392(7), 593–600. PMID:21657980. 1.3.4

[Bartel, 2004] Bartel, D. (2004). Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, 116(2), 281–297. PMID:14744438. 1.3.2

[Berman et al., 2007] Berman, H., Henrick, K., Nakamura, H., & Markley, J. (2007). The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic Acids Res*, 35, D301–D303. PMID:17142228. 6

[Bernstein et al., 2001] Bernstein, E., Caudy, A., Hammond, S., & Hannon, G. (2001). Role for a bidentate ribonuclease in the initiation step of rna interference. *Nature*, 409(6818), 363–366. PMID:11201747. 1.3.5, 6

[Berriman et al., 2009] Berriman, M., Haas, B., LoVerde, P., Wilson, R., Dillon, G., Cerqueira, G., Mashiyama, S., Al-Lazikani, B., Andrade, L., Ashton, P., Aslett, M., Bartholomeu, D., Blandin, G., Caffrey, C., Coghlan, A., Coulson, R., Day, T., Delcher, A., DeMarco, R., Djikeng, A., Eyre, T., Gamble, J., Ghedin, E., Gu, Y., Hertz-Fowler, C., Hirai, H., Hirai, Y., Houston, R., Ivens, A., Johnston, D., Lacerda, D., Macedo, C., McVeigh, P., Ning, Z., Oliveira, G., Overington, J., Parkhill, J., Pertea, M., Pierce, R., Protasio, A., Quail, M., Rajandream, M., Rogers, J., Sajid, M., Salzberg, S., Stanke, M., Tivey, A., White, O., Williams, D., Wortman, J., Wu, W., Zamanian, M., Zerlotini, A., Fraser-Liggett, C., Barrell, B., & El-Sayed, N. (2009). The genome of the blood fluke schistosoma mansoni. *Nature*, 460(7253), 352–358. PMID:19606141. 4.2.2

[Birikh et al., 1997] Birikh, K., Heaton, P., & Eckstein, F. (1997). The structure, function and application of the hammerhead ribozyme. *Eur J Biochem*, 245(1), 1–16. PMID:9128718. 1.3.4

[Boesler et al., 2011] Boesler, C., Kruse, J., Söderbom, F., & Hammann, C. (2011). Sequence and generation of mature ribosomal rna transcripts in dictyostelium discoideum. *J Biol Chem*, 286(20), 17693–17703. PMID:21454536. 3.3, 4.5

[Boguski et al., 1993] Boguski, M., Lowe, T., & Tolstoshev, C. (1993). dbest–database for ëxpressed sequence tags". *Nat Genet.*, 4(4), 332–333. PMID:8401577. 2.1

[Branch & Robertson, 1984] Branch, A. & Robertson, H. (1984). A replication cycle for viroids and other small infectious rna's. *Science*, 223(4635), 450–455. PMID:6197756. 1.11, 4.2.2

[Brazma et al., 1998] Brazma, A., Jonassen, I., Eidhammer, I., & Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5, 279–305. PMID:9672833. 1.4

[Breaker, 2012] Breaker, R. (2012). Riboswitches and the rna world. *Cold Spring Harb Perspect Biol*, 4(2), pii:a003566. PMID:21106649. 1.3.4

[Buskiewicz & Burke, 2012] Buskiewicz, I. & Burke, J. (2012). Folding of the hammerhead ribozyme: Pyrrolo-cytosine fluorescence separates core folding from global folding and reveals a ph-dependent conformational change. *RNA*, 2. PMID:22274955. 1.3.4

[Buzayan et al., 1986a] Buzayan, J., Gerlach, W., & Bruening, G. (1986a). Non-enzymatic cleavage and ligation of rnas complementary to a plant virus satellite rna. *Nature*, 323, 349–353. doi:10.1038/323349a0. 1.3.4

[Buzayan et al., 1986b] Buzayan, J., Gerlach, W., Bruening, G., Keese, P., & Gould, A. (1986b). Nucleotide sequence of satellite tobacco ringspot virus rna and its relationship to multimeric forms. *Virology*, 151(2), 186–199. PMID:18640637. 1.3.4, 1.7

[Byun & Han, 2009] Byun, Y. & Han, K. (2009). Pseudoviewer3: generating planar drawings of large-scale rna structures with pseudoknots. *Bioinformatics*, 25(11), 1435–1437. PMID:19369500. 2.4.13, 3.14

[Canny et al., 2004] Canny, M., Jucker, F., Kellogg, E., Khvorova, A., Jayasena, S., & Pardi, A. (2004). Fast cleavage kinetics of a natural hammerhead ribozyme. *J Am Chem Soc*, 126(35), 10848–10849. PMID:15339162. 1.3.4

[Cech et al., 1981] Cech, T., Zaug, A., & Grabowski, P. (1981). In vitro splicing of the ribosomal rna precursor of tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, 27, 487–496. PMID:6101203. 1.3.4

[Cegan et al., 2012] Cegan, R., Vyskot, B., Kejnovsky, E., Kubat, Z., Blavet, H., Safář, J., Doležel, J., Blavet, N., & Hobza, R. (2012). Genomic diversity in two related plant species with and without sex chromosomes - silene latifolia and s. vulgaris. *PLoS One*, 7(2), e31898. PMID:22393373. 6

[Chen et al., 2009] Chen, J., Gong, B., Bevilacqua, P., Carey, P., & Golden, B. (2009). A catalytic metal ion interacts with the cleavage site g.u wobble in the hdv ribozyme. *Biochemistry*, 48(7), 1498–1507. PMID:19178151. 1.3.4

[Chen et al., 2010] Chen, J., Yajima, R., Chadalavada, D., Chase, E., Bevilacqua, P., & Golden, B. (2010). A 1.9 a crystal structure of the hdv ribozyme precleavage suggests both lewis acid and general acid mechanisms contribute to phosphodiester cleavage. *Biochemistry*, 49(31), 6508–6518. PMID:20677830. 1.9

[Chi et al., 2008] Chi, Y., Martick, M., Lares, M., Kim, R., Scott, W., & Kim, S. (2008). Capturing hammerhead ribozyme structures in action by modulating general base catalysis. *PLoS Biol*, 6(9), e234. PMID:18834200. 1.3.4

[Chung et al., 2007] Chung, Y., Lee, W., Tang, C., & Lu, C. (2007). Re-music: a tool for multiple sequence alignment with regular expression constraints. *Nucleic Acids Res*, 35, W639–644. PMID:17488842. 2.4.17

[Cochrane & Strobel, 2008] Cochrane, J. & Strobel, S. (2008). Catalytic strategies of self-cleaving ribozymes. *Acc Chem Res*, 41(8), 1027–1035. PMID:18652494. 1.3.4, 1.3.4, 6

[Cock et al., 2010] Cock, P., Fields, C., Goto, N., Heuer, M., & Rice, P. (2010). The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Res*, 38(6), 1767–1771. PMID:20015970. 1.9.1

[Cornish-Bowden, 1985] Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res*, 13, 3021–3030. PMID:2582368. 1.3.1, 2.4.12, 2.4.12, 2.4.17

[Cremisi et al., 1992] Cremisi, F., Scarabino, D., Carluccio, M., Salvadori, P., & Barsacchi, G. (1992). A newt ribozyme: a catalytic activity in search of a function. *Proc Natl Acad Sci U S A*, 89(5), 1651–1655. PMID:1542657. 2.4.17

[Crick, 1966] Crick, F. (1966). Codon–anticodon pairing: the wobble hypothesis. *J Mol Biol*, 19(2), 548–555. PMID:5969078. 1.3

[Crick, 1968] Crick, F. (1968). The origin of the genetic code. *J Mol Biol*, 38(3), 367–379. PMID:4887876. 1.3.2

[Crick, 1970] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227, 561–563. PMID 4913914. 1.1, 1.2, 1.3.2

[Crombach & Hogeweg, 2011] Crombach, A. & Hogeweg, P. (2011). Is rna-dependent rna polymerase essential for transposon control? *BMC Syst Biol*, 5, 104. PMID:21714914. 4.5

[Cruz & Westhof, 2011] Cruz, J. & Westhof, E. (2011). Sequence-based identification of 3d structural modules in rna with rmdetect. *Nature Methods*, 8(6), 513–519. doi:10.1038/nmeth.1603. 1.4

[Daròs & Flores, 1995] Daròs, J. & Flores, R. (1995). Identification of a retroviroid-like element from plants. *Proc Natl Acad Sci U S A*, 92(15), 6856–6860. PMID:7542779. 1.3.4

[de la Peña et al., 2003] de la Peña, M., Gago, S., & Flores, R. (2003). Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity. *EMBO J*, 22(20), 5561–5570. PMID:14532128. 1.3.4

[de la Peña & Garcia-Robles, 2010a] de la Peña, M. & Garcia-Robles, I. (2010a). Intronic hammerhead ribozymes are ultraconserved in the human genome. *EMBO Rep*, 11(9), 711–716. PMID:20651741. 3.2.4, 4.2, 4.2.2

[de la Peña & Garcia-Robles, 2010b] de la Peña, M. & Garcia-Robles, I. (2010b). Ubiquitous presence of the hammerhead ribozyme motif along the tree of life. *RNA*, 16, 1943–1950. PMID:20705646. 1.4, 3.2.4, 4.2, 4.2.2, 6

[de la Peña et al., 1999] de la Peña, M., Navarro, B., & Flores, R. (1999). Mapping the molecular determinant of pathogenicity in a hammerhead viroid: a tetraloop within the in vivo branched rna conformation. *Proc Natl Acad Sci U S A*, 96(17), 9960–9965. PMID:10449802. 5

[Deigan & Ferré-D'Amaré, 2011] Deigan, K. & Ferré-D'Amaré, A. (2011). Riboswitches: discovery of drugs that target bacterial gene-regulatory rnas. *Acc Chem Res*, 44(12), 1329–1338. PMID:21615107. 1.3.4

[Deininger & Batzer, 2002] Deininger, P. & Batzer, M. (2002). Mammalian retroelements. *Genome Res*, 12(10), 1455–1465. PMID:12368238. 1.3.3

[Dirks et al., 2004] Dirks, R., Lin, M., Winfree, E., & Pierce, N. (2004). Paradigms for computational nucleic acid design. *Nucleic Acids Res*, 32(4), 1392–1403. PMID:14990744. 1.3.1

[Dsouza et al., 1997] Dsouza, M., Larsen, N., & Overbeek, R. (1997). Searching for patterns in genomic data. *Trends Genet*, 13(12), 497–498. PMID:9433140. 2.3, 2.4.12, 4.1.1

[Dufour et al., 2009] Dufour, D., de la Peña, M., Gago, S., Flores, R., & Gallego, J. (2009). Structure-function analysis of the ribozymes of chrysanthemum chlorotic mottle viroid: a loop-loop interaction motif conserved in most natural hammerheads. *Nucleic Acids Res*, 37(2), 368–381. PMID:19043070. 1.3.4

[Dunoyer et al., 2010] Dunoyer, P., Schott, G., Himber, C., Meyer, D., Takeda, A., Carrington, J., & Voinnet, O. (2010). Small rna duplexes function as mobile silencing signals between plant cells. *Science*, 328(5980), 912–916. PMID:20413458. 1.3.5

[Eddy, 2005a] Eddy, S. (2005a). Rnabob 2.1. ftp://selab.janelia.org/pub/software/rnabob/. 2.3, 2.4.12

[Eddy, 2005b] Eddy, S. (2005b). Squid 1.9g: A function library for sequence analysis. http://selab.wustl.edu/cgi-bin/selab.pl?mode=software. 2.3, 2.4.12

[Edgar, 2004] Edgar, R. (2004). Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5, 113. PMID:15318951. 2.4.7, 3.1.1, 6

[Eichinger et al., 2005] Eichinger, L., Pachebat, J., Glockner, G., Rajandream, M., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., Tunggal, B., Kummerfeld, S., Madera, M., Konfortov, B., Rivero, F., Bankier, A., Lehmann, R., Hamlin, N., Davies, R., Gaudet, P., Fey, P., Pilcher, K., Chen, G., Saunders, D., Sodergren, E., Davis, P., Kerhornou, A., Nie, X., Hall, N., Anjard, C., Hemphill, L., Bason, N., Farbrother, P., Desany, B., Just, E., Morio, T., Rost, R., Churcher, C., Cooper, J., Haydock, S., van Driessche, N., Cronin, A., Goodhead, I., Muzny, D., T, M., Pain, A., Lu, M., Harper, D., Lindsay, R., Hauser, H., James, K., Quiles, M., Babu, M. M., Saito, T., Buchrieser, C., Wardroper, A., Felder, M., Thangavelu, M., Johnson, D., Knights, A., Loulseged, H., Mungall, K., Oliver, K., Price, C., Quail, M., Urushihara, H., Hernandez, J., Rabbinowitsch, E., Steffen, D., Sanders, M., Ma, J., Kohara, Y., Sharp, S., Simmonds, M., Spiegler, S., Tivey, A., Sugano, S., White, B., Walker, D., Woodward, J., Winckler, T., Tanaka, Y., Shaulsky, G., Schleicher, M., Weinstock, G., Rosenthal, A., Cox, E., Chisholm, R., Gibbs, R., Loomis, W., Platzer, M., Kay, R., Williams, J., Dear, P., Noegel, A., Barrell, B., & Kuspa, A. (2005). The genome of the social amoeba dictyostelium discoideum. *Nature*, 435(7038), 43–57. PMID:15875012. 1.3.3, 6

[El-Murr et al., 2012] El-Murr, N., Maurel, M., Rihova, M., Vergne, J., Hervé, G., Kato, M., & Kawamura, K. (2012). Behavior of a hammerhead ribozyme in aqueous solution at medium to high temperatures. *Naturwissenschaften*, 99(9), 731–738. PMID:22915317. 4.2.2

[Elena et al., 1991] Elena, S., Dopazo, J., Flores, R., Diener, T., & Moya, A. (1991). Phylogeny of viroids, viroidlike satellite rnas, and the viroidlike domain of hepatitis delta virus rna. *Proc Natl Acad Sci U S A*, 88(13), 5631–5634. PMID:1712103. 1.3.4

[Epstein & Gall, 1987] Epstein, L. & Gall, J. (1987). Self-cleaving transcripts of satellite dna from the newt. *Cell*, 48(3), 535–543. PMID:2433049. 1.3.4, 2.4.17, 3.2.4

[Fedor, 1999] Fedor, M. (1999). Tertiary structure stabilization promotes hairpin ribozyme ligation. *Biochemistry*, 38(34), 11040–11050. PMID:10460159. 1.3.4, 1.3.4

[Fedor, 2009] Fedor, M. (2009). Comparative enzymology and structural biology of rna self-cleavage. *Annu Rev Biophys*, 38, 271–299. PMID:19416070. 1.3.4, 6

[Feldstein et al., 1989] Feldstein, P., Buzayan, J., & Bruening, G. (1989). Two sequences participating in the autolytic processing of satellite tobacco ringspot virus complementary rna. *Gene*, 82(1), 53–61. PMID:2583519. 1.3.4

[Ferbeyre et al., 2000] Ferbeyre, G., Bourdeau, V., Pageau, M., Miramontes, P., & Cedergren, R. (2000). Distribution of hammerhead and hammerhead-like rna motifs through the genbank. *Genome Res*, 10(7), 1011–1019. PMID:10899150. 1.3.4, 2.4.17

[Ferbeyre et al., 1998] Ferbeyre, G., Smith, J., & Cedergren, R. (1998). Schistosome satellite dna encodes active hammerhead ribozymes. *Mol Cell Biol*, 18(7), 3880–3888. PMID:9632772. 1.3.4, 2.4.17, 3.2.4, 3.18

[Ferré-D'Amaré et al., 1998] Ferré-D'Amaré, A., Zhou, K., & Doudna, J. (1998). Crystal structure of a hepatitis delta virus ribozyme. *Nature*, 395(6702), 567–574. PMID:9783582. 1.3.4, 1.3.4

[Finn et al., 2007] Finn, R., Stalker, J., Jackson, D., Kulesha, E., Clements, J., & Pettett, R. (2007). Proserver: a simple, extensible perl das server. *Bioinformatics*, 23(12), 1568–1570. PMID:17237073. 2.4.16, 2.4, 2.4.16, 6

[Fire et al., 1998] Fire, A., Xu, S., Montgomery, M., Kostas, S., Driver, S., & Mello, C. (1998). Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391(6669), 806–811. PMID:9486653. 1.3.5

[Forster & Symons, 1987] Forster, A. & Symons, R. (1987). Self-cleavage of plus and minus rnas of a virusoid and a structural model for the active sites. *Cell*, 49, 211–220. DOI:10.1016/0092-8674(87)90562-9. 1.3.4

[Gent et al., 2010] Gent, J., Lamm, A., Pavelec, D., Maniar, J., Parameswaran, P., Tao, L., Kennedy, S., & Fire, A. (2010). Distinct phases of sirna synthesis in an endogenous rnai pathway in c. elegans soma. *Mol Cell*, 37(5), 679–689. PMID:20116306. 1.3.5

[Gent et al., 2009] Gent, J., Schvarzstein, M., Villeneuve, A., Gu, S., Jantsch, V., Fire, A., & Baudrimont, A. (2009). A caenorhabditis elegans rna-directed rna polymerase in sperm development and endogenous rna interference. *Genetics*, 183(4), 1297–1314. PMID:19805814. 1.3.5

[Gilbert, 1986] Gilbert, W. (1986). Origin of life: The rna world. *Nature*, 319(6055), 618. doi:10.1038/319618a0. 1.1

[Giliberti et al., 2012] Giliberti, J., O'Donnell, S., Etten, W. V., & Janssen, G. (2012). A 5'-terminal phosphate is required for stable ternary complex formation and translation of leaderless mrna in escherichia coli. *RNA*. PMID:22291205. 1.3.4

[Glöckner & Heidel, 2009] Glöckner, G. & Heidel, A. (2009). Centromere sequence and dynamics in dictyostelium discoideum. *Nucleic Acids Res*, 37(6), 1809–1816. PMID:19179372. 4.5, 6

[Glöckner et al., 2001] Glöckner, G., Szafranski, K., Winckler, T., Dingermann, T., Quail, M., Cox, E., Eichinger, L., Noegel, A., & Rosenthal, A. (2001). The complex repeats of dictyostelium discoideum. *Genome Res*, 11(4), 585–594. PMID:11282973. 1.3.3

[Golden, 2011] Golden, B. (2011). Two distinct catalytic strategies in the hepatitis delta virus ribozyme cleavage reaction. *Biochemistry*, 50(40), 9424–9433. PMID:22003985. 1.3.4, 1.3.4, 1.9

[Green et al., 1993] Green, B., Pabón-Peña, L., Graham, T., Peach, S., Coats, S., & Epstein, L. (1993). Conserved sequence and functional domains in satellite 2 from three families of salamanders. *Mol Biol Evol*, 10(4), 732–750. PMID:8355598. 2.4.17

[Griffiths-Jones et al., 2003] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., & Eddy, S. (2003). Rfam: an rna family database. *Nucleic Acids Res*, 31(1), 439–441. PMID:12520045. 1.4

[Grillo et al., 2003] Grillo, G., Licciulli, F., Liuni, S., Sbisà, E., & Pesole, G. (2003). Patsearch: A program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res*, 31(13), 3608–3612. PMID:12824377. 2.4.12

[Grivna et al., 2006] Grivna, S., Beyret, E., Wang, Z., & Lin, H. (2006). A novel class of small rnas in mouse spermatogenic cells. *Genes Dev*, 20(13), 1709–1714. PMID:16766680. 1.3.2

[Gräf et al., 2005] Gräf, S., Teune, J., Strothmann, D., Kurtz, S., & Steger, G. (2005). *A Computational Approach to Search for Non-Coding RNAs in Large Genomic Data*, volume 17, (pp. 57–74). Springer-Verlag. 1.4, 1.6, 2.1, 2.4.12, 2.4.17

[Gräf, 2005] Gräf, S. A. (2005). *Strukturbasierte Beschreibung und Suche von RNA-Familien in genomischen Sequenzdaten*. Doktorarbeit, Heinrich-Heine-Universität Düsseldorf. 2.4.14

[Guerrier-Takada et al., 1983] Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., & Altman, S. (1983). The rna moiety of ribonuclease p is the catalytic subunit of the enzyme. *Cell*, 35, 849–857. PMID:6197186. 1.3.4

[Hall, 1999] Hall, T. (1999). Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. *Nucl. Acids. Sym. Ser.*, 41, 95–98. http://www.mbio.ncsu.edu/bioedit/bioedit.html. 2.4.6

[Hammann & Lilley, 2002] Hammann, C. & Lilley, D. (2002). Folding and activity of the hammerhead ribozyme. *ChemBioChem*, 3, 690–700. PMID: 12203967. 1.3.4

[Hammann et al., 2012] Hammann, C., Luptak, A., Perreault, J., & de la Peña, M. (2012). The ubiquitous hammerhead ribozyme. *RNA*, 18(5), 871–885. PMID:22454536. 1.3.4, 4.2, 4.2.2

[Hammann & Steger, 2012] Hammann, C. & Steger, G. (2012). Viroid-specific small rna in plant disease. *RNA Biol*, 9(6), ePub. PMID:22617880. 1.3.4

[Hammann & Westhof, 2007] Hammann, C. & Westhof, E. (2007). Searching genomes for ribozymes and riboswitches. *Genome Biol*, 8(4), 210. PMID:17472738. 1.3.4

[Haseloff & Gerlach, 1988] Haseloff, J. & Gerlach, W. (1988). Simple rna enzymes with new and highly specific endoribonuclease activities. *Nature*, 334(6183), 585–591. PMID:2457170. 1.3.4

[Hertel et al., 2007] Hertel, J., Hofacker, I., & Stadler, P. (2007). Snoreport: Computational identification of snornas with unknown targets. *Bioinformatics*, 24, 158–164. PMID:17895272. 1.3.2

[Hertel et al., 1992] Hertel, K., Pardi, A., Uhlenbeck, O., Koizumi, M., Ohtsuka, E., Uesugi, S., Cedergren, R., Eckstein, F., Gerlach, W., & Hodgson, R. (1992). Numbering system for the hammerhead. *Nucleic Acids Research*, 20(12), 3252. PMCID: PMC312468. 1.3.4, 1.5, 2.7, 3.1.1

[Hertz & Stormo, 1999] Hertz, G. & Stormo, G. (1999). Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15, 563–577. PMID:10487864. 1.5, 6

[Hinas et al., 2007] Hinas, A., Reimegård, J., Wagner, E., Nellen, W., Ambros, V., & Söderbom, F. (2007). The small rna repertoire of dictyostelium discoideum and its regulation by components of the rnai pathway. *Nucleic Acids Res*, 35(20), 6714–6726. PMID:17916577. 1.3.2, 1.3.3, 1.3.5, 3.3, 4.5

[Hizver et al., 2001] Hizver, J., Rozenberg, H., Frolow, F., Rabinovich, D., & Shakked, Z. (2001). Dna bending by an adenine–thymine tract and its role in gene regulation. *Proc Natl Acad Sci U S A*, 98(15), 8490–8495. PMID:11438706. 1.2, 1.1

[Hofacker et al., 1994] Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, S., Tacker, M., & Schuster, P. (1994). *Fast folding and comparison of RNA secondary structures*, volume 125, (pp. 167–188). Springer-Verlag. Monatshefte für Chemie. 1.3.1, 2.4.13, 4.1.1

[Hoffgaard et al., Prep] Hoffgaard, F., Seehafer, C., Hammann, C., & Hamacher, K. (Prep). Information theory reveals distinct evolutionary patterns in ribozymes. unpublished. (document), 1.3.4, 3.2.4, 3.2.4, 3.25, 4.2.2, 6

[Hong et al., 1999] Hong, Y., Ontiveros, S., Chen, C., & Strauss, W. (1999). A new structure for the murine xist gene and its relationship to chromosome choice/counting during x-chromosome inactivation. *Proc Natl Acad Sci U S A*, 96(12), 6829–6834. PMID:10359798. 1.3.2

[Hubbard et al., 2009] Hubbard, T., Aken, B., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graef, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski1, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., & Flicek, P. (2009). Ensembl 2009. *Nucleic Acids Research*, 37. D690-D697. 1

[Hutchins et al., 1986] Hutchins, C., Rathjen, P., Forster, A., & Symons, R. (1986). Self-cleavage of plus and minus rna transcripts of avocado sunblotch viroid. *Nucleic Acids Res*, 14(9), 3627–3640. PMID:3714492. 1.3.4, 6

[Hüttenhofer et al., 2001] Hüttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J., & Brosius, J. (2001). Rnomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger rnas in mouse. *EMBO Journal*, 20(11), 2943–2953. PMID:11387227. 1.3.2, 6

[Hüttenhofer & Vogel, 2006] Hüttenhofer, A. & Vogel, J. (2006). Experimental approaches to identify non-coding rnas. *Nucleic Acids Research*, 34(2), 635–646. PMID:16436800. 1.3.2

[Jimenez et al., 2011] Jimenez, R., Delwart, E., & Lupták, A. (2011). Structure-based search reveals hammerhead ribozymes in the human microbiome. *J Biol Chem*, 286(10), 7737–7743. PMID:21257745. 1.3.4, 3.2.4, 4.2.2, 5

[Jurka et al., 2005] Jurka, J., Kapitonov, V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110(1-4), 462–467. PMID:16093699. 2.4.18

[Jöchl et al., 2008] Jöchl, C., Rederstorff, M., Hertel, J., Stadler, P., Hofacker, I., Schrettl, M., Haas, H., & Hüttenhofer, A. (2008). Aspergillus fumigatus suggests a novel mechanism for regulation of protein synthesis. *Nucleic Acids Research*, 36(8), 2677–2689. PMID:18346967. 1.3.2, 1.4

[Kalweit et al., 2011] Kalweit, A., Przybilski, R., Seehafer, C., de la Peña, M., & Hammann, C. (2011). *Characterization of Hammerhead Ribozyme reactions*, chapter 2. Springer Verlag. (document), 1.4, 6

[Kaper et al., 1988] Kaper, J., Tousignant, M., & Steger, G. (1988). Nucleotide sequence predicts circularity and self-cleavage of 300-ribonucleotide satellite of arabis mosaic virus. *Biochem Biophys Res Commun*, 154(1), 318–325. PMID:3395334. 1.7

[Kartha, 1967] Kartha, G. (1967). Tertiary structure of ribonuclease. *Nature*, 214(5085), 234. PMID:6034236. 4.2.2

[Kennedy et al., 2008] Kennedy, R., Lladser, M., Yarus, M., & Knight, R. (2008). Information, probability, and the abundance of the simplest rna active sites. *Front Biosci*, 13, 6060–6071. PMID:18508643. 4.4

[Kennell et al., 1995] Kennell, J., Saville, B., Mohr, S., Kuiper, M., Sabourin, J., Collins, R., & Lambowitz, A. (1995). The vs catalytic rna replicates by reverse transcription as a satellite of a retroplasmid. *Genes Dev*, 9(3), 294–303. PMID:7532606. 1.3.4

[Khvorova et al., 2003] Khvorova, A., Lescoute, A., Westhof, E., & Jayasena, S. (2003). Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. *Nat Struct Biol*, 10(9), 708–712. PMID:12881719. 1.3.4

[Kim & Kim, 2007] Kim, Y. & Kim, V. (2007). Processing of intronic microrna. *EMBO J*, 26(3), 775–783. PMID:17255951. 4.2.2

[Klein & Ferré-D'Amaré, 2006] Klein, D. & Ferré-D'Amaré, A. (2006). Structural basis of glms ribozyme activation by glucosamine-6-phosphate. *Science*, 313(5794), 1752–1756. PMID:16990543. 1.3.4

[Klenk et al., 1997] Klenk, H., Clayton, R., Tomb, J., White, O., Nelson, K., Ketchum, K., Dodson, R., Gwinn, M., Hickey, E., Peterson, J., Richardson, D., Kerlavage, A., Graham, D., Kyrpides, N., Quackenbush, R. F. J., Lee, N., Sutton, G., Gill, S., Kirkness, E., Dougherty, B., McKenney, K., Adams, M., Loftus, B., Peterson, S., Reich, C., McNeil, L., Badger, J., Glodek, A., Zhou, L., Overbeek, R., Gocayne, J., Weidman, J., McDonald, L., Utterback, T., Cotton, M., Spriggs, T., Artiach, P., Kaine, B., Sykes, S., Sadow, P., D'Andrea, K., Bowman, C., Fujii, C., Garland, S., Mason, T., Olsen, G., Fraser, C., Smith, H., Woese, C., & Venter, J. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon archaeoglobus fulgidus. *Nature*, 390(6658), 364–370. PMID:9389475. 4.2.2

[Kozomara & Griffiths-Jones, 2011] Kozomara, A. & Griffiths-Jones, S. (2011). mirbase: integrating microrna annotation and deep-sequencing data. *Nucleic Acids Res*, 39, D152–D157. PMID:21037258. 3.3

[Kruger et al., 1982] Kruger, K., Grabowski, P., Zaug, A., Sands, J., Gottschling, D., & Cech, T. (1982). Self-splicing rna: autoexcision and autocyclization of the ribosomal rna intervening sequence of tetrahymena. *Cell*, 31(1), 147–157. PMID:6297745. 1.1, 1.3.4

[Kuhlmann et al., 2005] Kuhlmann, M., Borisova, B., Kaller, M., Larsson, P., Stach, D., Na, J., Eichinger, L., Lyko, F., Ambros, V., Söderbom, F., Hammann, C., & Nellen, W. (2005). Silencing of retrotransposons in dictyostelium by dna methylation and rnai. *Nucleic Acids Res*, 33(19), 6405–6417. PMID:16282589. 1.3.5, 4.5, 6

[Kuo et al., 1988] Kuo, M., Sharmeen, L., Dinter-Gottlieb, G., & Taylor, J. (1988). Characterization of self-cleaving rna sequences on the genome and antigenome of human hepatitis delta virus. *J Virol*, 62(12), 4439–4444. PMID:3184270. 1.3.4

[Laferriere et al., 1994] Laferriere, A., Gautheret, D., & Cedergren, R. (1994). An rna pattern matching program with enhanced performance and portability. *Computer Applications in the Biosciences*, 10(2), 211–212. DOI:10.1093/bioinformatics/10.2.211. 2.4.12

[Lafontaine et al., 2002] Lafontaine, D., Norman, D., & Lilley, D. (2002). The global structure of the vs ribozyme. *EMBO J*, 21, 2461–2471. PMID:12006498. 1.3.4

[Lai, 1995] Lai, M. (1995). The molecular biology of hepatitis delta virus. *Annu Rev Biochem*, 64, 259–286. PMID:7574482. 1.3.4

[Langmead et al., 2009] Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3), R25. PMID:19261174. 2.4.18

[Lassmann & Sonnhammer, 2005] Lassmann, T. & Sonnhammer, E. (2005). Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6, 298. PMID:16343337. 2.4.7, 6

[Lawrence et al., 1993] Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., & Wootton, J. (1993). Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262, 208–214. PMID:8211139. 1.5

[Lee et al., 2008] Lee, K., Backer, P. D., Aono, T., Liu, C., Suzuki, S., Suzuki, T., Kaneko, T., Yamada, M., Tabata, S., Kupfer, D., Najar, F., Wiley, G., Roe, B., Binnewies, T., Ussery, D., D'Haeze, W., Herder, J., Gevers, D., Vereecke, D., Holsters, M., & Oyaizu, H. (2008). The genome of the versatile nitrogen fixer azorhizobium caulinodans ors571. *BMC Genomics*, 9, 271. PMID:18522759. 4.2.2

[Leontis et al., 2002] Leontis, N., Stombaugh, J., & Westhof, E. (2002). The non-watson-crick base pairs and their associated isostericity matrices. *Nucleic Acids Res*, 30(16), 3497–3531. PMID:12177293. 1.2

[Leontis & Westhof, 2001] Leontis, N. & Westhof, E. (2001). Geometric nomenclature and classification of rna base pairs. *RNA*, 7(4), 499–512. PMID:11345429. 1.3, 1.2, 1.6

[Levenshtein, 1966] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10, 707–710. www.freearchive.org. 6

[Li & Ding, 2005] Li, H. & Ding, S. (2005). Antiviral silencing in animals. *FEBS Lett*, 579(26), 5965–5973. PMID:16154568. 1.3.2, 1.3.5

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16), 2078–2079. PMID:19505943. 1.9.7, 1.9.8, 2.4.18

[Lilley, 2004] Lilley, D. (2004). The varkud satellite ribozyme. *RNA*, 10(2), 151–158. PMID:14730013. 1.3.4

[Lipfert et al., 2008] Lipfert, J., Ouellet, J., Norman, D., Doniach, S., & Lilley, D. (2008). The complete vs ribozyme in solution studied by small-angle x-ray scattering. *Structure*, 16(9), 1357–1367. PMID:18786398. 1.3.4, 1.10

[Lowe & Eddy, 1997] Lowe, T. & Eddy, S. (1997). trnascan-se: a program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Res*, 25, 955–964. PMID: 9023104. 1.4

[Lu et al., 2006] Lu, C., Kulkarni, K., Souret, F., MuthuValliappan, R., Tej, S., Poethig, R., Henderson, I., Jacobsen, S., Wang, W., Green, P., & Meyers, B. (2006). Micrornas and other small rnas enriched in the arabidopsis rna-dependent rna polymerase-2 mutant. *Genome Res*, 16(10), 1276–1288. PMID:16954541. 4.5

[Luzi et al., 1997] Luzi, E., Eckstein, F., & Barsacchi, G. (1997). The newt ribozyme is part of a riboprotein complex. *Proc Natl Acad Sci U S A*, 94(18), 9711–9716. PMID:9275189. 1.3.4

[Lévesque & Perreault, 2012] Lévesque, M. & Perreault, J. (2012). Target-induced sofa-hdv ribozyme. *Methods Mol Biol*, 848, 369–384. PMID:22315081. 1.3.4

[Macke, 2001] Macke, T. (2001). Rnamotif, an rna secondary structure definition and search algorithm. *Nucleic Acids Res*, 29, 4724–4735. PMID:11713323. 1.4

[Maida & Masutomi, 2011] Maida, Y. & Masutomi, K. (2011). Rna-dependent rna polymerases in rna silencing. *Biol Chem*, 392(4), 299–304. PMID:21294682. 1.3.5, 1.3.5, 4.5, 6

[Maida et al., 2009] Maida, Y., Yasukawa, M., Furuuchi, M., Lassmann, T., Possemato, R., Okamoto, N., Kasim, V., Hayashizaki, Y., Hahn, W., & Masutomi, K. (2009). An rna-dependent rna polymerase formed by tert and the rmrp rna. *Nature*, 461(7261), 230–235. PMID:19701182. 1.3.5

[Makeyev & Bamford, 2002] Makeyev, E. & Bamford, D. (2002). Cellular rna-dependent rna polymerase involved in posttranscriptional gene silencing has two distinct activity modes. *Mol Cell*, 10(6), 1417–1427. PMID:12504016. 1.3.5

[Manber & Myers, 1993] Manber, U. & Myers, E. (1993). Suffix arrays: a new method for on-line string searches. *SIAM J. Comput*, 33(5), 935–948. http://goanna.cs.rmit.edu.au/∼e76763/pub/mm93-joc.pdf. 2.4.12

[Maniar & Fire, 2011] Maniar, J. & Fire, A. (2011). Ego-1, a c. elegans rdrp, modulates gene expression via production of mrna-templated short antisense rnas. *Curr Biol*, 21(6), 449–459. PMID:21396820. 1.3.5

[Markham & Zuker, 2005] Markham, N. & Zuker, M. (2005). Dinamelt web server for nucleic acid melting prediction. *Nucleic Acids Res*, 33, 577–581. PMID:15980540. 1.7

[Markham & Zuker, 2008] Markham, N. & Zuker, M. (2008). *UNAFold: software for nucleic acid folding and hybridization*, volume 2, chapter 1, (pp. 3–31). Humana Press. PMID:18712296. 2.3, 2.4.13

[Martens et al., 2002] Martens, H., Novotny, J., Oberstrass, J., Steck, T., Postlethwait, P., & Nellen, W. (2002). Rnai in dictyostelium: the role of rna-directed rna polymerases and double-stranded rnase. *Mol Biol Cell*, 13(2), 445–453. PMID:11854403. 1.3.5

[Martick et al., 2008] Martick, M., Horan, H., Noller, H., & Scott, W. (2008). A discontinuous hammerhead ribozyme embedded in a mammalian messenger rna. *Nature*, 454(7206), 899–902. PMID:18615019. 1.3.4, 1.3.4, 1.3.4, 3.2.4, 3.6

[Martick & Scott, 2006] Martick, M. & Scott, W. (2006). Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell*, 126, 309–320. PMID:16859740. 1.3, 1.3.4, 1.6

[Martin, 2011] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1). 2.4.18

[Mathews et al., 1999] Mathews, D., Sabina, J., Zuker, M., & Turner, D. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *J Mol Biol*, 288(5), 911–940. PMID:10329189. 2.4.13

[Mathias et al., 1991] Mathias, S., Kazazian, A. S. H. J., Boeke, J., & Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. *Science*, 254(5039), 1808–1810. PMID:1722352. 1.3.3

[McGuire & Galagan, 2008] McGuire, A. & Galagan, J. (2008). Conserved secondary structures in aspergillus. *PLoS One*, 3(7), e2812. PMID:18665251. 1.3.2

[Moissiard et al., 2007] Moissiard, G., Parizotto, E., Himber, C., & Voinnet, O. (2007). Transitivity in arabidopsis can be primed, requires the redundant action of the antiviral dicer-like 4 and dicer-like 2, and is compromised by viral-encoded suppressor proteins. *RNA*, 13(8), 1268–1278. PMID:17592042. 1.3.5

[Molnar et al., 2010] Molnar, A., Melnyk, C., Bassett, A., Hardcastle, T., Dunn, R., & Baulcombe, D. (2010). Small silencing rnas in plants are mobile and direct epigenetic modification in recipient cells. *Science*, 328(5980), 872–875. PMID:20413459. 1.3.5

[Moss, 1996] Moss, G. (1996). Basic terminology of stereochemistry (iupac recommendations 1996). *Pure Appl. Chem.*, 68(12), 2193–2222. doi:10.1351/pac199668122193. 1.3.1

[Motamedi et al., 2004] Motamedi, M., Verdel, A., Colmenares, S., Gerber, S., Gygi, S., & Moazed, D. (2004). Two rnai complexes, rits and rdrc, physically interact and localize to noncoding centromeric rnas. *Cell*, 119(6), 789–802. PMID:15607976. 1.3.5

[Nawrocki et al., 2009] Nawrocki, E., Kolbe, D., & Eddy, S. (2009). Infernal 1.0: inference of rna alignments. *Bioinformatics*, 25(10), 1335–1337. PMID:19307242. 4.1.1

[Nissen et al., 2000] Nissen, P., Hansen, J., Ban, N., Moore, P., & Steitz, T. (2000). The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289(5481), 920–930. PMID:10937990. 1.3.2, 1.3.4

[Notredame et al., 2000] Notredame, C., Higgins, D., & Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1), 205–217. PMID:10964570. 3.1.1

[Osborne et al., 2005] Osborne, E., Schaak, J., & Derose, V. (2005). Characterization of a native hammerhead ribozyme derived from schistosomes. *RNA*, 11(2), 187–196. PMID:15659358. 1.3.4

[Pak & Fire, 2007] Pak, J. & Fire, A. (2007). Distinct populations of primary and secondary effectors during rnai in c. elegans. *Science*, 315(5809), 241–244. PMID:17124291. 1.3.5

[Pearson, 1990] Pearson, W. (1990). Rapid and sensitive sequence comparison with fastp and fasta. *Methods Enzymol*, 183, 63–98. PMID:2156132. 2.4.17

[Penedo et al., 2004] Penedo, J., Wilson, T., Jayasena, S., Khvorova, A., & Lilley, D. (2004). Folding of the natural hammerhead ribozyme is enhanced by interaction of auxiliary elements. *RNA*, 10(5), 880–888. PMID:15100442. 1.3.4

[Perreault et al., 2011] Perreault, J., Weinberg, Z., Roth, A., Popescu, O., Chartrand, P., Ferbeyre, G., & Breaker, R. (2011). Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. *PLoS Comput Biol*, 7(5), e1002031. PMID:21573207. 1.3.4, 1.3.4, 1.3.4, 1.3.4, 3.2.4, 3.2.4, 3.2.4, 3.2.4, 4.1.1, 4.2, 4.2.2, 6

[Pesole et al., 2000] Pesole, G., Liuni, S., & D'Souza, M. (2000). Patsearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, 16(5), 439–450. PMID:10871266. 2.4.12

[Pevzner & Sze, 2000] Pevzner, P. & Sze, S. (2000). Combinatorial approaches to finding subtle signals in dna sequences. *Proc Int Conf Intell Syst Mol Biol*, 8, 269–278. PMID:10977088. 1.4, 1.5, 6

[Pinheiro et al., 2012] Pinheiro, V., Taylor, A., Cozens, C., Abramov, M., Renders, M., Zhang, S., Chaput, J., Wengel, J., Peak-Chew, S., McLaughlin, S., Herdewijn, P., & Holliger, P. (2012). Synthetic genetic polymers capable of heredity and evolution. *Science*, 336(6079), 341–344. PMID:22517858. 1.1

[Prody et al., 1986] Prody, G., Bakos, J., Buzayan, J., Schneider, I., & Bruening, G. (1986). Autolytic processing of dimeric plant virus. *Science*, 231(4745), 1577–1580. PMID:17833317. 1.3.4, 1.3.4, 1.3.4, 1.3.4, 6

[Przybilski et al., 2005] Przybilski, R., Gräf, S., Lescoute, A., Nellen, W., Westhof, E., Steger, G., & Hammann, C. (2005). Functional hammerhead ribozymes naturally encoded in the genome of arabidopsis thaliana. *The Plant Cell*, 17, 1877–1885. PMID: 15937227. 1.3.4, 3.2.4, 3.2.4, 3.2.4, 4.1.1, 4.2, 4.2.2, 6

[Przybilski & Hammann, 2006] Przybilski, R. & Hammann, C. (2006). The hammerhead ribozyme structure brought in line. *Chembiochem*, 7(11), 1641–1644. PMID:16991176. 1.3.4

[Przybilski & Hammann, 2007] Przybilski, R. & Hammann, C. (2007). The tolerance to exchanges of the watson crick base pair in the hammerhead ribozyme core is determined by surrounding elements. *RNA*, 13(10), 1625–1630. PMID:17666711. 1.3.4, 4.2.2

[Rajasekaran et al., 2005] Rajasekaran, S., Balla, S., Huang, C.-H., Thapar, V., Gryk, M., Maciejewski, M., & Schiller, M. (2005). High-performance exact algorithms for motif search. *Journal of Clinical Monitoring and Computing*, 19, 319–328. DOI: 10.1007/s10877-005-0677-y. 1.4

[Rastogi et al., 1996] Rastogi, T., Beattie, T., Olive, J., & Collins, R. (1996). A long-range pseudoknot is required for activity of the neurospora vs ribozyme. *EMBO J*, 15, 2820–2825. PMID:8654379. 1.3.4

[Reeder & Giegerich, 2004] Reeder, J. & Giegerich, R. (2004). Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5, 104. PMID:15294028. 1.4, 2.4.13

[Reeder & Giegerich, 2009] Reeder, J. & Giegerich, R. (2009). *RNA secondary structure analysis using the RNAshapes package*, chapter 12. Curr Protoc Bioinformatics. PMID:19496058. 4.1.2

[Reeder et al., 2007] Reeder, J., Reeder, J., & Giegerich, R. (2007). Locomotif: from graphical motif description to rna motif search. *Bioinformatics*, 23(13), i392–i400. PMID:17646322. 1.4, 4.1.1

[Reiter et al., 2011] Reiter, N., Chan, C., & Mondragón, A. (2011). Emerging structural themes in large rna molecules. *Curr Opin Struct Biol*, 21(3), 319–326. PMID:21474301. 1.3.1

[Richterich, 1998] Richterich, P. (1998). Estimation of errors in "raw"dna sequences: a validation study. *Genome Res*, 8(3), 251–259. PMID:9521928. 6

[Rivas et al., 2001] Rivas, E., Klein, R., & Eddy, T. J. S. (2001). Computational identification of noncoding rnas in e. coli by comparative genomics. *Curr Biol*, 11(17), 1369–1373. PMID:11553332. 1.4

[Robinson et al., 2011] Robinson, J., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G., & Mesirov, J. (2011). Integrative genomics viewer. *Nat Biotechnol*, 29(1), 24–26. PMID:21221095. 2.4.18

[Rocheleau & Pelchat, 2006] Rocheleau, L. & Pelchat, M. (2006). The subviral rna database: a toolbox for viroids, the hepatitis delta virus and satellite rnas research. *BMC Microbiol*, 6, 24. PMID:16519798. 1.7, 2.4.17, 3.1.1, 3.2.4, 4.2, 6

[Rojas et al., 2000] Rojas, A., Vazquez-Tello, A., Ferbeyre, G., Venanzetti, F., Bachmann, L., Paquin, B., Sbordoni, V., & Cedergren, R. (2000). Hammerhead-mediated processing of satellite pdo500 family transcripts from dolichopoda cave crickets. *Nucleic Acids Res*, 28(20), 4037–4043. PMID:11024185. 1.3.4, 2.4.17, 3.2.4

[Roychowdhury-Saha et al., 2011] Roychowdhury-Saha, M., Roychowdhury, S., & Burke, D. (2011). Conformational heterogeneity and the determinants of tertiary stabilization in the hammerhead ribozyme from dolichopoda cave crickets. *RNA Biol*, 8(5). PMID:21712651. 1.3.4, 4.2.2

[Rubino et al., 1990] Rubino, L., Tousignant, M., Steger, G., & Kaper, J. (1990). Nucleotide sequence and structural analysis of two satellite rnas associated with chicory yellow mottle virus. *J Gen Virol*, 71(9), 1897–1903. PMID:1698918. 1.7

[Ruffner et al., 1990] Ruffner, D., Stormo, G., & Uhlenbeck, O. (1990). Sequence requirements of the hammerhead rna self-cleavage reaction. *Biochemistry*, 29(47), 10695–10702. PMID:1703005. 1.3.4

[Ruminski et al., 2011] Ruminski, D., Webb, C., Riccitelli, N., & Lupták, A. (2011). Processing and translation initiation of non-long terminal repeat retrotransposons by hepatitis delta virus (hdv)-like self-cleaving ribozymes. *J Biol Chem*, 286(48), 41286–41295. PMID:21994949. 1.3.4, 3.2.2, 4.2.3, 6

[Rupert & Ferré-D'Amaré, 2001] Rupert, P. & Ferré-D'Amaré, A. (2001). Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature*, 410(6830), 780–786. PMID:11298439. 1.3.4, 1.3.4, 1.3.4, 1.8

[Saitou & Nei, 1987] Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4), 406–425. PMID:3447015. 2.4.7, 2.4.7, 2.4.7

[Salehi-Ashtiani et al., 2006] Salehi-Ashtiani, K., Lupták, A., Litovchick, A., & Szostak, J. (2006). A genomewide search for ribozymes reveals an hdv-like sequence in the human cpeb3 gene. *Science*, 313(5794), 1788–1792. PMID:16990549. 6

[Salehi-Ashtiani & Szostak, 2001] Salehi-Ashtiani, K. & Szostak, J. (2001). In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature*, 414(6859), 82–84. PMID:11689947. 1.3.4, 4.2.2, 6

[Saville & Collins, 1990] Saville, B. & Collins, R. (1990). A site-specific self-cleavage reaction performed by a novel rna in neurospora mitochondria. *Cell*, 61(4), 685–696. PMID:2160856. 1.3.4, 1.3.4, 6

[Schwartz et al., 2003] Schwartz, S., Kent, W., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D., & Miller, W. (2003). Human-mouse alignments with blastz. *Genome Res*, 13(1), 103–107. PMID:12529312. 2.4.12

[Scott et al., 1995] Scott, W., Finch, J., & Klug, A. (1995). The crystal structure of an all-rna hammerhead ribozyme: a proposed mechanism for rna catalytic cleavage. *Cell*, 81(7), 991–1002. PMID:7541315. 1.6

[Seehafer et al., 2012] Seehafer, C., Kalweit, A., & Hammann, C. (2012). Genomisch codierte hammerhead-ribozyme. *Biospektrum*, 18(5), 484–486. DOI:10.1007/s12268-012-0217-5. (document), 1.11, 3.2.4, 3.21

[Seehafer et al., 2011] Seehafer, C., Kalweit, A., Steger, G., Gräf, S., & Hammann, C. (2011). From alpaca to zebrafish: Hammerhead ribozymes wherever you look. *RNA*, 17(1), 21–26. PMID:21081661. (document), 3.10, 3.2.4, 3.2.4, 3.4, 3.5, 3.2.4, 3.23, 4.2, 4.2.1, 4.2.2, 6

[Seemann & Hartig, 2011] Seemann, I. & Hartig, J. (2011). Artificial ribozyme-based regulators of gene expression. *Synlett*, 11, 1486–1494. DOI:10.1055/s-0030-1260583. 1.3.4, 1.4

[Sewer et al., 2005] Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M., Tuschl, T., van Nimwegen, E., & Zavolan, M. (2005). Identification of clustered micrornas using an ab initio prediction method. *BMC Bioinformatics*, 6(267), Epub. PMID:16274478. 2.4.12

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 623–656. http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html. 6

[Shippy et al., 1999] Shippy, R., Lockner, R., Farnsworth, M., & Hampel, A. (1999). The hairpin ribozyme. discovery, mechanism, and development for gene therapy. *Mol Biotechnol*, 12(1), 117–129. PMID:10554775. 1.3.4

[Siebert & Backofen, 2005] Siebert, S. & Backofen, R. (2005). Marna: multiple alignment and consensus structure prediction of rnas based on sequence structure comparisons. *Bioinformatics*, 21(16), 3352–3359. PMID:15972285. 2.4.17

[Sing et al., 2005] Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20), 3940–3941. PMID:16096348. 1.8, 3.1.2

[Sleutels et al., 2002] Sleutels, F., Zwart, R., & Barlow, D. (2002). The non-coding air rna is required for silencing autosomal imprinted genes. *Nature*, 415(6873), 810–813. PMID:11845212. 1.3.2

[Slotkin & Martienssen, 2007] Slotkin, R. & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*, 8(4), 272–285. PMID:17363976. 4.5

[Sokal & Michener, 1958] Sokal, R. & Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409–1438. 2.4.7, 6

[Stage-Zimmermann & Uhlenbeck, 1998] Stage-Zimmermann, T. & Uhlenbeck, O. (1998). Hammerhead ribozyme kinetics. *RNA*, 4(8), 875–889. PMID:9701280. 3.23, 6

[Staple & Butcher, 2005] Staple, D. & Butcher, S. (2005). Pseudoknots: Rna structures with diverse functions. *PLoS Biol*, 3(6), e213. PMID:15941360. 6

[Studier & Keppler, 1988] Studier, J. & Keppler, K. (1988). A note on the neighbor-joining algorithm of saitou and nei. *Mol Biol Evol*, 5(6), 729–731. PMID:3221794. 2.4.7

[Suydam et al., 2010] Suydam, I., Levandoski, S., & Strobel, S. (2010). Catalytic importance of a protonated adenosine in the hairpin ribozyme active site. *Biochemistry*, 49(17), 3723–3732. PMID:20373826. 1.3.4

[Tabler & Tsagris, 2004] Tabler, M. & Tsagris, M. (2004). Viroids: Petite rna pathogens with distinguished talents. *Trends Plant Sci*, 9, 339–348. PMID:15231279. 1.3.4, 3.2.4, 4.2

[Tamura & Holbrook, 2002] Tamura, M. & Holbrook, S. (2002). Sequence and structural conservation in rna ribose zippers. *J Mol Biol*, 320(3), 455–474. PMID:12096903. 6

[Team, 2008] Team, R. D. C. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 2.4.10

[Temin, 1964] Temin, H. (1964). The participation of dna in rous sarcoma virus production. *Virology*, 23, 486–494. PMID:14204701. 1.1

[Thompson et al., 1994] Thompson, J., Higgins, D., & Gibson, T. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22), 4673–4680. PMID:7984417. 2.4.7, 2.4.7, 3.1.1

[Tinoco & Bustamante, 1999] Tinoco, I. & Bustamante, C. (1999). How rna folds. *J Mol Biol*, 293(2), 271–281. PMID:10550208. 3.2.4

[Tinoco et al., 2004] Tinoco, J. I., Collin, D., & Li, P. (2004). The effect of force on thermodynamics and kinetics: unfolding single rna molecules. *Biochem Soc Trans*, 32, 757–760. PMID:15494007. 1.7

[Tomari et al., 2004] Tomari, Y., Du, T., Haley, B., Schwarz, D., Bennett, R., Cook, H., Koppetsch, B., Theurkauf, W., & Zamore, P. (2004). Risc assembly defects in the drosophila rnai mutant armitage. *Cell*, 116(6), 831–841. PMID:15035985. 1.3.2

[Trapnell et al., 2010] Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., & Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5), 511–515. PMID:20436464. 2.4.18

[Uhlenbeck, 1987] Uhlenbeck, O. (1987). A small catalytic oligoribonucleotide. *Nature*, 328(6131), 596–600. PMID:2441261. 1.3.4, 4.2.2

[Vanin, 1985] Vanin, E. (1985). Processed pseudogenes: characteristics and evolution. *Annu Rev Genet*, 19, 253–272. PMID:3909943. 1.3.3

[Vastenhouw et al., 2003] Vastenhouw, N., Fischer, S., Robert, V., Thijssen, K., Fraser, A., Kamath, R., Ahringer, J., & Plasterk, R. (2003). A genome-wide screen identifies 27 genes involved in transposon silencing in c. elegans. *Curr Biol*, 13(15), 1311–1316. PMID:12906791. 1.3.3

[Vazquez-Tello et al., 2002] Vazquez-Tello, A., Castán, P., Moreno, R., Smith, J., Berenguer, J., & Cedergren, R. (2002). Efficient trans-cleavage by the schistosoma mansoni smalpha1 hammerhead ribozyme in the extreme thermophile thermus thermophilus. *Nucleic Acids Res*, 30(7), 1606–1612. PMID:11917021. 4.2.2

[Verdaguer & Ferrer-Orta, 2012] Verdaguer, N. & Ferrer-Orta, C. (2012). Conformational changes in motif d of rdrps as fidelity determinant. *Structure*, 20(9), 1448–1450. PMID:22958639. 1.3.5

[Verdel et al., 2004] Verdel, A., Jia, S., Gerber, S., Sugiyama, T., Gygi, S., Grewal, S., & Moazed, D. (2004). Rnai-mediated targeting of heterochromatin by the rits complex. *Science*, 303(5658), 672–676. PMID:14704433. 1.3.5

[Viladoms et al., 2011] Viladoms, J., Scott, L., & Fedor, M. (2011). An active-site guanine participates in glms ribozyme catalysis in its protonated state. *J Am Chem Soc*, 133(45), 18388–18396. PMID:21936556. 1.3.4

[Wahl et al., 2009] Wahl, M., Will, C., & Lührmann, R. (2009). The spliceosome: design principles of a dynamic rnp machine. *Cell*, 136(4), 701–718. PMID:19239890. 1.3.2

[Wassenegger et al., 1994] Wassenegger, M., Heimes, S., Riedel, L., & Sänger, H. (1994). Rna-directed de novo methylation of genomic sequences in plants. *Cell*, 76(3), 567–576. PMID:8313476. 1.3.5

[Watson & Crick, 1953] Watson, J. & Crick, F. (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361), 964–967. PMID:13063483. 1.1, 1.2

[Webb & Lupták, 2011] Webb, C. & Lupták, A. (2011). Hdv-like self-cleaving ribozymes. *RNA Biol*, 8(5), Epub. PMID:21734469. 4.2.3

[Webb et al., 2009] Webb, C., Riccitelli, N., Ruminski, D., & Lupták, A. (2009). Widespread occurrence of self-cleaving ribozymes. *Science*, 326, 953–989. DOI: 10.1126/science.1178084. 1.3.4, 2.4.12, 2.12, 3.2.2, 4.2.3, 6

[Wiegand et al., 2011] Wiegand, S., Kruse, J., Gronemann, S., & Hammann, C. (2011). Efficient generation of gene knockout plasmids for dictyostelium discoideum using one-step cloning. *Genomics*, 97(5), 321–325. PMID:21316445. 2.4.18, 6

[Wiegand et al., Prep] Wiegand, S., Seehafer, C., Hofmann, P., Schmith, A., Winckler, T., Földesi, B., Boesler, B., Nellen, W., Reimegård, J., Käller, M., Hällman, J., Emanuelsson, O., Avesson, L., Söderbom, F., & Hammann, C. (Prep). The two rna-dependent rna polymerases rrpa and rrpc have distinct functions in the generation of small rnas and in the regulation of retrotransposons in dictyostelium discoideum. unpublished. (document), 1.3.5, 3.33, 4.5, 6

[Wilson & Lilley, 2009] Wilson, T. & Lilley, D. (2009). The evolution of ribozyme chemistry. *Science*, 323(5920), 1436–1438. PMID:19286542. 1.3.4

[Wilson & Lilley, 2011] Wilson, T. & Lilley, D. (2011). Do the hairpin and vs ribozymes share a common catalytic mechanism based on general acid-base catalysis? a critical assessment of available experimental data. *RNA*, 17(2), 213–221. PMID:21173201. 1.3.4, 1.3.4, 1.10, 4.2.4

[Winkler et al., 2004] Winkler, W., Nahvi, A., Roth, A., Collins, J., & Breaker, R. (2004). Control of gene expression by a natural metabolite-responsive ribozyme. *Nature*, 428(6980), 281–286. PMID:15029187. 1.3.4

[Wu et al., 1989] Wu, H., Lin, Y., Lin, F., Makino, S., Chang, M., & Lai, M. (1989). Human hepatitis delta virus rna subfragments contain an autocleavage activity. *Proc Natl Acad Sci U S A*, 86(6), 1831–1835. PMID:2648383. 1.3.4

[Xayaphoummine et al., 2005] Xayaphoummine, A., Bucher, T., & Isambert, H. (2005). Kinefold web server for rna/dna folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.*, 33(Web Server issue), W605–10. PMID:15980546. 2.4.13

[Yao et al., 2006] Yao, Z., Weinberg, Z., & Ruzzo, W. (2006). Cmfinder–a covariance model based rna motif finding algorithm. *Bioinformatics*, 22(4), 445–452. PMID:16357030. 1.4

[Zanger, 2005] Zanger, S. (2005). *Vergleich von Suchprogrammen für RNA-Motive*. Diplomarbeit, Heinrich-Heine-Universität Düsseldorf. 2.4.12, 4.1.1

[Zhang & Ruvkun, 2012] Zhang, C. & Ruvkun, G. (2012). New insights into sirna amplification and rnai. *RNA Biol*, 9(8), Epub. PMID:22858672. 1.3.5

[Zhang & Epstein, 1996] Zhang, Y. & Epstein, L. (1996). Cloning and characterization of extended hammerheads from a diverse set of caudate amphibians. *Gene*, 172(2), 183–190. PMID:8682301. 1.3.4

[Zharkikh & Li, 1992] Zharkikh, A. & Li, W. (1992). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. i. four taxa with a molecular clock. *Mol Biol Evol*, 9(6), 1119–1147. PMID:1435238. 2.4.7

[Zuker, 2003] Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31, 3406–3415. PMID:12824337. 2.3, 2.4.13

[Zuker & Stiegler, 1981] Zuker, M. & Stiegler, P. (1981). Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1), 133–148. PMID:6163133. 1.3.1

Lebenslauf

Der Lebenslauf ist in der Online Version aus Gründen des Datenschutzes modifiziert worden.

Persönliche Informationen:

Carsten Seehafer

Ausbildungsdaten:

Promotion:	04.2011 – 03.2012	Wissenschaftlicher Mitarbeiter an der TU Darmstadt (AG Hammann)
	02.2009 – 03.2011	Wissenschaftlicher Mitarbeiter an der Universität Kassel (AG Hammann)
Studium:		Abschluss: Master of Science
	10.2005 – 09.2008	Master in Bioinformatik an der Eberhard Karls Universität Tübingen
		Abschluss: Bachelor of Science
	10.2002 – 08.2005	Bachelor in Bioinformatik an der Freien Universität Berlin

Weitere Informationen:

Stipendium:	11.2011 – 12.2011	EMBO (short term fellowship)